Deep Learning in Mammography and Breast Histology, an Overview and Future Trends

Hamidinekoo, Azam; Denton, Erika R. E.; Rampun, Yambu Andrik; Honnor, Kate; Zwiggelaar, Reyer

Published in:
Medical Image Analysis

DOI:
10.1016/j.media.2018.03.006

Publication date:
2018

Citation for published version (APA):
Hamidinekoo, A., Denton, E. R. E., Rampun, Y. A., Honnor, K., & Zwiggelaar, R. (2018). Deep Learning in Mammography and Breast Histology, an Overview and Future Trends. *Medical Image Analysis*, *47*, 45-67. https://doi.org/10.1016/j.media.2018.03.006

Deep Learning in Mammography and Breast Histology, an Overview and Future Trends

Azam Hamidinekoo, Erika Denton, Andrik Rampun, Kate Honnor, Reyer Zwiggelaar

PII: S1361-8415(18)30090-2
DOI: 10.1016/j.media.2018.03.006
Reference: MEDIMA 1351

To appear in: *Medical Image Analysis*

Received date: 26 July 2017
Revised date: 3 January 2018
Accepted date: 14 March 2018

Please cite this article as: Azam Hamidinekoo, Erika Denton, Andrik Rampun, Kate Honnor, Reyer Zwiggelaar, Deep Learning in Mammography and Breast Histology, an Overview and Future Trends, *Medical Image Analysis* (2018), doi: 10.1016/j.media.2018.03.006

Deep Learning in Mammography and Breast Histology, an Overview and Future Trends

Azam Hamidinekoo[a,*], Erika Denton[b], Andrik Rampun[c], Kate Honnor[d], Reyer Zwiggelaar[a]

[a]*Department of Computer Science, Aberystwyth University, United Kingdom*
[b]*Department of Radiology, Norfolk and Norwich University Hospital, United Kingdom*
[c]*School of Computing, Ulster University, Coleraine, Northern Ireland, United Kingdom*
[d]*Department of Histopathology/Cytopathology, Norfolk and Norwich University Hospital, United Kingdom*

Abstract

Recent improvements in biomedical image analysis using deep learning based neural networks could be exploited to enhance the performance of Computer Aided Diagnosis (CAD) systems. Considering the importance of breast cancer worldwide and the promising results reported by deep learning based methods in breast imaging, an overview of the recent state-of-the-art deep learning based CAD systems developed for mammography and breast histopathology images is presented. In this study, the relationship between mammography and histopathology phenotypes is described, which takes biological aspects into account. We propose a computer based breast cancer modelling approach: the Mammography-Histology-Phenotype-Linking-Model, which develops a mapping of features/phenotypes between mammographic abnormalities and their histopathological representation. Challenges are discussed along with the potential contribution of such a system to clinical decision making and treatment management.

Keywords: Mammography, Breast Histopathology, Computer Aided Diagnosis, Deep Learning

*Corresponding author; Department of Computer Science, Aberystwyth University, UK
Email addresses: azh2@aber.ac.uk (Azam Hamidinekoo), erika.denton@nnuh.nhs.uk (Erika Denton), y.rampun@ulster.ac.uk (Andrik Rampun), kate.honnor@nnuh.nhs.uk (Kate Honnor), rrz@aber.ac.uk (Reyer Zwiggelaar)

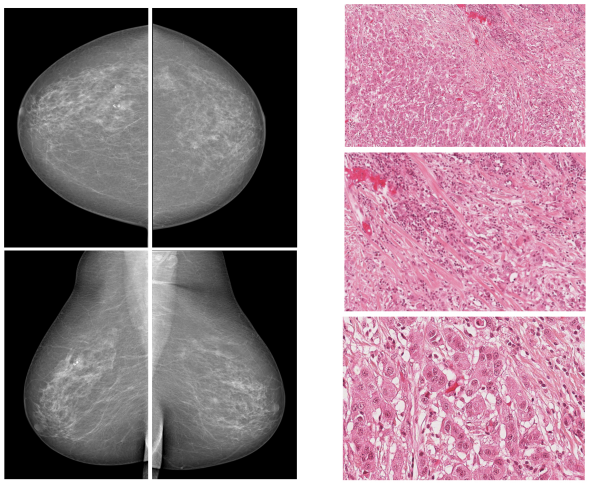
1. Introduction

1.1. Breast cancer

Breast cancer is the most frequently diagnosed cancer (National-Health-Service (2016); American-Cancer-Society (2016)) and accounts for 25.2% of the total cancer related deaths among women followed by colorectal (9.2%), lung (8.7%), cervix (7.9%), and stomach cancers (4.8%) according to the International Agency for Research on Cancer, WHO [1] (Stewart & Kleihues (2014)). The assessment process for breast screening follows a triple assessment model: appropriate imaging (i.e. mammography as a primary imaging modality for lesion visualisation and finding early changes in breast tissue) plus clinical assessment and, where indicated, needle biopsy (i.e. H&E [2] stained histology) (Breast-Cancer-Biopsy (2016)). Typical examples of mammographic and H&E histological images of breast tissue, as the two commonly used imaging modalities, are shown in Figure 1 and are the focus of this paper.

Among the women who undergo mammographic screening, about 10% are recalled for additional evaluation. Among these, 8 to 10% will have suspicious abnormal findings which warrant undergoing breast biopsy (Neal et al. (2010)). In the United States, approximately 15-30% referred for biopsy are found to have malignant abnormalities and in European trials, this ranges from 30% to 75% (Kopans (1992)). Although effective, this process is a trade-off between sensitivity (84%) and specificity (91%) which leads to a number of unnecessary biopsies (Elmore et al. (2009)). The impact of unnecessary biopsy and the downstream diagnostic burden includes increased anxiety, morbidity and stress for the women concerned and increased health care costs. Nevertheless, biopsy is currently considered the only way to confirm the presence of cancer (Elmore et al. (2009)). Therefore, there is a clear need to develop a specific discrimination

[1] World Health Organisation
[2] Hematoxylin and Eosin



(a) (b)

Figure 1: Two breast imaging modalities: (a) mammography images from the INBreast dataset (Moreira et al. (2012)), Craniocaudal (CC) and Mediolateral Oblique (MLO) views (left - right sides) shown in the first and the second row respectively; (b) breast histology images from the MITOS-ATYPIA-14 (2016) dataset, showing from top to bottom: 10 HPF, 20 HPF and 40 HPF (HPF stands for High Power Field which indicates magnified areas).

model or criteria, like the "Stavros Criteria" in ultrasound, which determines whether ultrasound could help accurately distinguish benign solid breast nodules from indeterminate or malignant nodules and whether this distinction could be specific enough to reduce the need for biopsy (Stavros et al. (1995)). In mammography, an equivalent model or criteria could indicate benign abnormalities and reduce the need for further biopsies.

1.2. Conventional CAD systems

In order to assist radiologists' interpretation, Computer Aided Diagnosis (CAD) systems and quantitative image analysis (QIA) techniques have been developed as an alternative to double reading, improving clinicians' accuracy and patient outcome. These systems are aimed at improved identification of subtle suspicious masses, calcifications, micro-calcifications and other abnormalities in mammograms (Oliver et al. (2010); He et al. (2015)). Meanwhile,

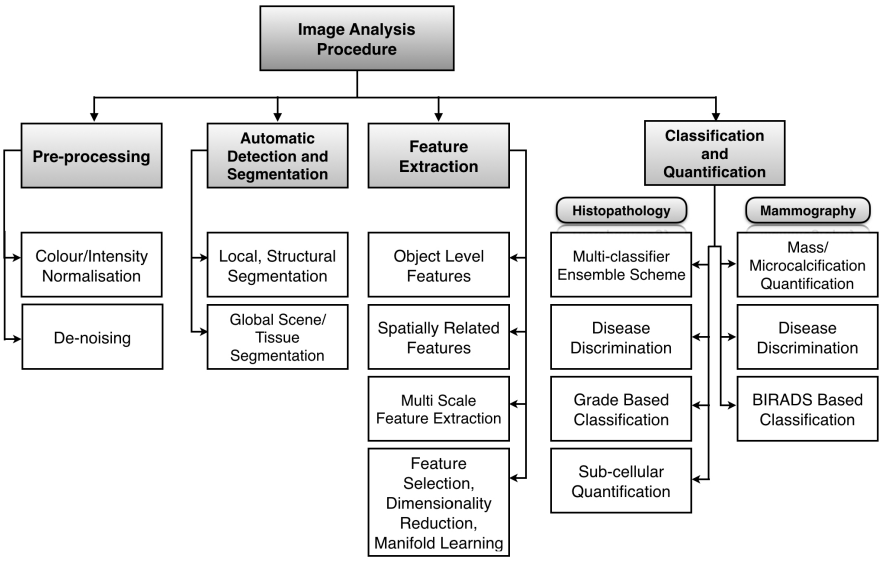


Figure 2: Image analysis procedures for mammography and histopathology image data.

histological CAD systems, provide another perspective on developing breast cancer models such as the identification of tumour regions, mitotic activity, nuclear atypia score, the epithelium-stroma and the tubule formation score along with identifying subtypes of breast cancer like IDC [3] or ILC [4] (Veta et al. (2014); Gurcan et al. (2009)). An overview of machine learning based image analysis aspects used in histopathology and mammography CAD systems is provided in Figure 2. Using conventional machine learning methods, various hand-designed descriptors (i.e. morphological, topological and textural features) based on prior knowledge and expert guidance have been developed for these CAD systems. Previous publications have described and compared such approaches for automatic detection and segmentation of abnormalities in mammographic images (Oliver et al. (2010, 2006); Giger et al. (2013); Boyer et al. (2009)). When dealing with breast histology; inherent complexities are

[3] Invasive Ductal Carcinoma
[4] Invasive Lobular Carcinoma

modelled via different algorithms to achieve specific tasks (Kowal et al. (2013); Irshad et al. (2014); Dundar et al. (2011); Kothari et al. (2013); Veillard et al. (2013)). These models and approaches have been evaluated on different breast databases including digital/digitised mammography and histology images.

The most significant weakness of conventional machine learning methods is the hand-engineered feature extraction step, which employs a combination of heuristic and mathematical descriptors. Subsequently, the extracted features are introduced into different classifiers to be categorised into the desired classes as expressed in Figure 2. This feature extraction step makes the learning algorithm more cumbersome since it mostly depends on the features extracted from the data and requires effort and sufficient interpreting knowledge due to the various geometrical and morphological structures. Reproducing results is not always easily achieved and the generic discrimination ability of the features used needs investigation.

1.3. Towards, deep learning based CAD systems

The benefits of conventional mammographic and histologic CAD systems in clinical practice have not been fully determined. There has been significant discussion on whether CAD is an effective tool at the current level of performance (Fenton et al. (2011); Giger (2014)). Still, more creative and predictive models need to be designed to improve the performance metrics, including accuracy, sensitivity, specificity, precision and recall rate to improve upon the current state-of-the art. A crucial step towards a new generation of machine learning approaches is enabling computers to learn the features as data representatives. These are expressed as low-level features such as margin and edge; middle-level features such as edge junctions and high level object parts (Zeiler et al. (2011)). Deep learning approaches - termed one of the significant breakthrough technologies of recent years by the MIT Technology Review (MIT-Technology-Review (2017))- has made headlines in producing semantic information due to its nature of adaptive learning from input data. Various deep learning structures have been developed for both supervised approaches (algo-

rithms that infer a function from input data with labelled responses) and unsupervised approaches (algorithms that draw inferences from input data without labelled responses). SAE [5] (Ng (2011)), DBN [6] and RBM [7] (Salakhutdinov & Hinton (2009)) are among popular architectures developed for unsupervised approaches. CNN [8] (LeCun et al. (2010, 1998)) and RNN [9] (Medsker & Jain (1999)) have become the technique of choice for supervised approaches. In recent years, a noticeable shift from conventional machine learning methods to deep learning based methods is seen in a wide variety of real world, especially medical, applications and several review papers have been published (Schmidhuber (2015); LeCun et al. (2015); Litjens et al. (2017)). Several open crowdsourced algorithmic analysis competitions have been announced to motivate the development of better techniques for cancer prognosis, detection, risk stratification, disease outcome prediction and survival. Recently held breast cancer mammography related competitions have been the Digital-Mammography-DREAM-Challenge (2017)) and UK-Breast-Cancer (2016). Some recent breast histopathology competitions include: ICPR2012 (2017)[10], AMIDA13 (2017)[11], MITOS-ATYPIA-14 (2016), CAMELYON16 (2016); CAMELYON17 (2017)[12] and TUPAC16 (2016)[13]. These competitions have influenced the evaluation of different methods to become more transparent and easier to compare. In most of these challenges, deep learning based approaches have shown the most promising performance.

In AI [14] technology, deep learning methods have multiple levels of representation learning which use raw data and discover the essential representations

[5] Sparse AutoEncoders
[6] Deep Belief Networks
[7] Restricted Boltzmann Machines
[8] Convolutional Neural Networks
[9] Recurrent Neural Networks
[10] International Conference on Pattern Recognition
[11] Assessment of Mitosis Detection Algorithms
[12] Cancer Metastasis Detection in Lymph Node
[13] Tumor Proliferation Assessment Challenge
[14] Artificial Intelligence

for detection or classification (LeCun et al. (2015)). These inherent representations and patterns are obtained through a hierarchical framework which is able to put features extracted from a low level (starting with raw data) and high level abstracts together using a non-linear approach. Such networks are able to improve themselves according to the input content variation and optimise the relationship between inputs and outputs via an iterative training process (Bengio (2009)). At the same time as the deep learning concepts were developed, a step-change in processing power through high performance GPUs [15] and open source frameworks/libraries developed on CUDA [16] (CUDA (2017)) or OpenCL [17] (OpenCL (2017)) platforms have made significant progress for the implementation of deep learning based methods. These open source frameworks and libraries provide the chance for optimised implementation of convolutions and other related functions. In addition, they facilitate the ability to perform a high number of computations at a relatively low costs through their massive parallel architectures.

1.4. Structure of the paper

This paper presents an overview of different deep learning based approaches used for mammography and breast histology and proposes a bridge between these two fields employing deep learning concepts. We have focused on mammography, since this is the most common modality used in breast screening, and H&E stained histology, since it is considered as the gold standard for final decision making.

The main aims of this paper are:

1. In Section 2, deep learning based models are introduced and their fundamental structures summarised.
2. Recent deep learning based approaches for mammographic and histopathologic image analysis are reviewed (covered in Sections 3 and 4, respec-

[15] Graphics Processing Units
[16] Compute Unified Device Architecture
[17] Open Computing Language

tively). Details of the models (e.g. datasets, architecture, etc.) are provided in separate tables.

3. Exploring the link between mammography and histology phenotypes from a biological point of view is reviewed in Section 5.
4. The future of deep learning in constructing a model linking mammographic and histologic features and phenotypes called "Mammography-Histology-Phenotype-Linking-Model" ($ML_{M<->H}$) is covered in Section 6.
5. Potential challenges to be considered in the development of $ML_{M<->H}$ are also discussed in Section 6.

1.4.1. Paper selection process

When selecting the papers, popular review papers (Veta et al. (2014); Gurcan et al. (2009); Oliver et al. (2010); Rangayyan et al. (2007); Doi (2007); He et al. (2015); Litjens et al. (2017)) were considered. Other papers citing them and publishing work on mammography or breast histology were also reviewed. Papers published by participants in breast cancer challenges were selected too. Google Scholar was searched using keywords: "*breast cancer, mammography, histopathology, CAD systems, deep learning, Convolutional Neural Network (CNN), linking map, phenotype*" and those related to breast cancer and deep learning were included in this review.

2. Deep Neural Networks

2.1. General architecture of deep neural networks

Various deep architectures have been derived from traditional feed-forward ANN [18]. An ANN consists of a cascade of trainable multi-stage layers inspired by the organisation of the animal visual cortex (LeCun et al. (2010)). There are sets of arrays called feature maps as the input and output of each layer. Each feature map in a specific layer represents particular features extracted at the locations of the associated input.

[18] Artificial Neural Network

Commonly used layers in deep learning based networks are:

Input layer. This loads input to feed the convolutional layers. Some transformations such as mean-subtraction, feature-scaling and effective data augmentation can be incorporated (Hamidinekoo et al. (2017)).

Convolutional layer. This tends to includes three stages of operational units (LeCun et al. (2010)):

- Convolutional filters: these compute the convolution result of the input feature map with trainable 2D discrete convolution filters and bias parameters. Each filter bank detects a particular feature at each location on the input map (LeCun et al. (2010); Schmidhuber (2015)).
- Pooling: this performs down-sampling for the spatial dimension of the input. This results in a reduced-resolution output feature map which is robust to small variations in the location of features in the previous layer. Additionally, it merges semantically similar features into one. There are a number of variations for pooling (i.e. maximum, average) (Krizhevsky & Hinton (2009)).
- Activation or non-linearity function: this is a non-linear element-wise operator that simulates excitability of neurons. Among various activation functions in deep learning, the Rectified Linear Unit (ReLU) has been shown to be efficient for image processing applications (Glorot et al. (2011); Dahl et al. (2013)).

Normalisation layer. This can be implemented at each spatial location across all feature maps of the same layer in order to acquire an improved description of the input. This way, non-uniformity of the scene illumination can be reduced which leads to better convergence by decorrelating the input dimensions (Dahl et al. (2013)).

Dropout regularisation layer. This can reduce over-fitting of the network and result in learning more robust features. The key idea is to randomly drop units along with the respective connections from the neural network during the training process to avoid too much co-adaptation of the units (Srivastava et al. (2014)).

Inner-product layers or fully connected layers. These treat their input as a simple vector and produce an output in the form of a single vector. In classification tasks, the last layers are sometimes fully-connected layers that are followed by logarithmic loss to be minimised. The exact merit of fully connected layers is still an open research question, but its effect in improving the performance has been reported (Krizhevsky et al. (2012)).

Constructing the architecture using these elements, a signal is propagated through active neurons from layer to layer. This signal is a linear combination of the input, learned weights and biases treated under a non-linearity function as:

$$signal = F_{nonlinear}(weights^T * input + bias) \quad (1)$$

Accordingly, in the forward direction the loss function (specifically defined for a task) is calculated. Optimisation of the calculated error is obtained using a form of stochastic gradient descent (LeCun et al. (1998)). Hence, coefficients of all filters in distinct layers are calculated and updated simultaneously during the learning process with the back-propagation method (LeCun et al. (2012)). Training is an iterative process involving multiple passes of the input data through the network until the model converges (LeCun et al. (2015); Schmidhuber (2015)).

Two of the most important types are Convolutional Neural Networks and AutoEncoders, which are described in Sections 2.2 and 2.3.

2.2. Convolutional Neural Networks (CNNs)

CNNs are the most successful type of deep learning model, especially for supervised learning applied to image based classification work. Litjens et al.

(2017) have published a comprehensive review on different image processing applications accomplished by CNNs. Like regular ANNs, CNNs are made up of several layers stacked on top of each other. However, unlike a regular Neural Network, the layers of a CNN have width, height and depth so that they are controllable by their depth and breath variations which enables them to share weights (Simonyan & Zisserman (2014)). A CNN can be trained by feeding it a suitable input. It is then able to compute parameters layer by layer and produce a final output. The objective of training is to minimise the difference between the predicted output and the actual output of the network. This error then flows backwards through the net by a back-propagation procedure and updates the parameter values. A typical CNN architecture is shown in Figure 3.

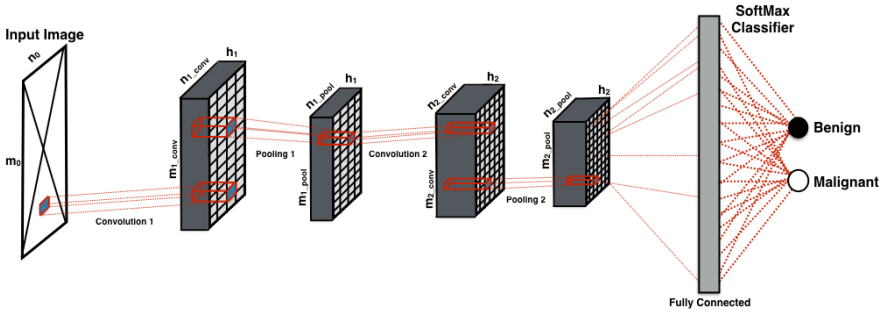


Figure 3: A typical Convolutional Neural Network architecture.

2.3. AutoEncoder

An AutoEncoder (AE) is a form of ANN, developed for unsupervised learning models (Bengio (2009)). An AE is able to learn generative representations from image data, typically with the purpose of reconstructing the input on the output layer by reducing the dimensionality space through the hidden layers. AEs have been widely used for segmentation and detection tasks in breast image analysis while CNNs are mostly used for the task of predicting a target value (i.e. classification). Architecturally, AEs are feed-forward, non-recurrent neural networks that consist of two parts: the encoder and the decoder. A schematic

architecture of an AE is shown in Figure 4. The objective of training is to minimise the reconstruction error which in the simplest form can be expressed as:

$$Loss(I,O) = ||I-O||^2 = ||I-F_{De}(W_{De}^T*(F_{En}(W_{En}^T*I+B_{En}))+B_{De})||^2 \quad (2)$$

I: Input image

O: Output image

F_{En}: Encoder element wise activation function

F_{De}: Decoder element wise activation function

W_{En}: Weight in Encoder

W_{De}: Weight in Decoder

B_{En}: Bias in Encoder

B_{De}: Bias in Decoder

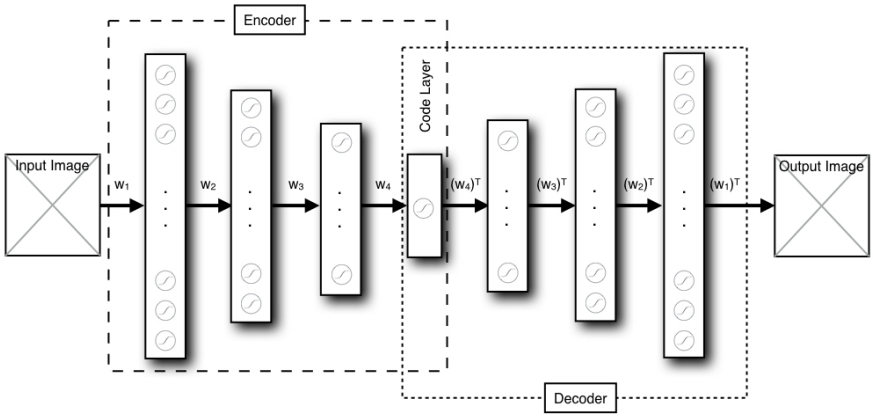


Figure 4: Schematic architecture of an AutoEncoder (AE).

2.4. Developed models

CNNs and AEs have several general advantages compared to conventional feed-forward neural networks such as: no dependency on designing hand-crafted features; reduced pre-processing analysis on input data; calculation of fewer connections and parameters; ability to pool similar features at the same location

and nearby locations due to the use of shared weights; and translation invariance (Donahue et al. (2014)). Moreover, the saturation issue and vanishing or exploding gradient of a layer, which are serious concerns for neural networks (Schmidhuber (2015)), can be addressed with careful choice of activation functions, careful weight initialisation and small learning rates during optimisation. Detailed description of these technical aspects are covered by Nair & Hinton (2010). For more detailed information about the mathematical concept of deep learning based architectures, the reader is suggested to consult LeCun et al. (2010).

The success or failure of a model depends on the aforementioned modifiable compartments of the learning system. Sections 3 and 4 will focus on the CNNs and AEs applied in mammography and histology image processing applications and how the proposed models have improved the state-of-the-art results for CAD systems in these two fields. There are several "standard" deep learning networks throughout this paper which are used in the developed models described in the later sections:

CifarNet (Krizhevsky & Hinton (2009)) has three convolution layers, three pooling layers, and one fully-connected layer. This CNN architecture has about 0.15 million free parameters.

AlexNet (Krizhevsky et al. (2012)) has five convolution layers, three pooling layers, and two fully-connected layers with approximately 61 million free parameters. It has halved the error rate in object recognition competitions and facilitated the rapid adoption of deep learning.

GoogLeNet proposed by Szegedy et al. (2015), is significantly more complex in structure and depth and introduced an "Inception" module that consisted of six convolution layers and one pooling layer which is responsible for concatenation of filters with different sizes and dimensions into a single new filter. Overall, GoogLeNet has two convolution layers, two pooling layers, and nine Inception layers leading to nearly 5 million free parameters.

VGGNet (Simonyan & Zisserman (2014)) showed the effect of the network depth on performance. It described 2 best versions: containing 16 and 19

convolution/fully-connected layers performing on 3×3 filter sizes with approximately 138 million free parameters in VGGNet 16.

2.5. Common challenges and proposed strategies in deep learning

In recent years, deep learning based methods have been considered the preferred approach for many medical imaging applications. However, in order to integrate them into application pipelines, some considerations should be taken into account. Comparison of various algorithmic methods is difficult since each research team has reported their results using their own dataset and evaluation metrics (Gurcan et al. (2009)). To address the issue of such variation, some data has been made publicly available. For example, in the Whole Slide Imaging Repository website[19], histopathology images and information for different organs is accessible. Of critical concern for supervised learning is the amount of annotated data available. To address this, some image data has been made publicly available in terms of competitions but still labelling them specifically is time-consuming, tedious and sometimes costly. The annotations should be done in a structured format to be usable by the larger community. A list of recent large datasets in mammography and breast histopathology are provided in Tables 1 and 2.

There are currently three major approaches for successfully employing supervised deep networks, which also address the issue of data availability, (i.e. for image classification via CNNs): i) training a network from scratch, ii) using off-the-shelf pre-trained network features and iii) using unsupervised networks and pre-training with supervised fine-tuning (Shin et al. (2016); Goodfellow et al. (2016)). Based on the reported results (Tajbakhsh et al. (2016)), CNNs are difficult to train from scratch for most medical images due to the small data sample sizes, variance in abnormality appearances and lack of rare or special cases. Transfer learning and fine-tuning in medical image analysis are two effective methods in which a network (i.e. a CNN model) is pre-trained on a

[19] `https://digitalpathologyassociation.org/whole-slide-imaging-repository`

natural image dataset or a different medical domain and then fine-tuned on the desired medical images. Thanks to some open source frameworks, like Caffe (Jia et al. (2014)), these pre-trained networks can simply be downloaded and directly applied to any medical image analysis.

Another solution to collecting a larger number of annotated image data is crowdsourcing (Albarqouni et al. (2016)). This technique allows for combining radiologists' or histopathologists' knowledge with non-experts to enable learning inputs from crowds as part of the network learning process. While the unlabelled data can never replace labelled data, using unlabelled data is also a supplement to the annotated data. Artificial data augmentation is another solution widely used for increasing the number of training cases (Hamidinekoo et al. (2017)). These issues are further addressed in the models covered in Sections 3 and 4.

3. Deep Learning in Mammographic Image Processing

3.1. Problem statement

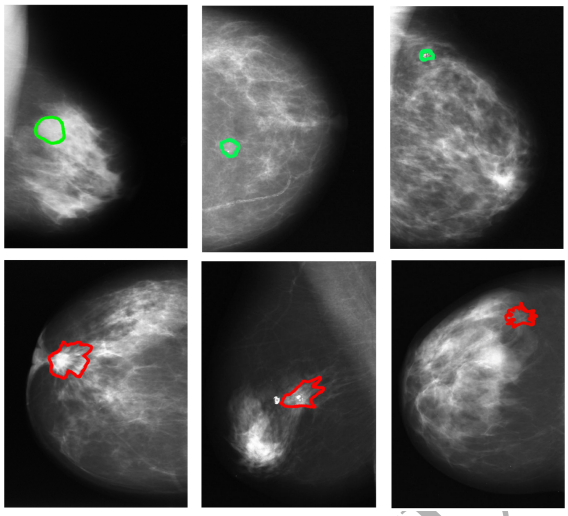
Mammograms reflect density variations in breast tissue composition due to different X-ray attenuation in breast tissue. Epithelium and stroma attenuate x-rays more than fat and thus appear radiopaque on mammograms while fat appears radiolucent (Tabár & Dean (2005)). Several studies have confirmed the relationship between breast cancer risk and mammographic parenchymal (texture) patterns assessed by percent mammographic density (Gastounioti et al. (2016)) (besides age, gender, gene mutations and family history factors). Breast cancer can appear in mammograms as: masses, architectural distortion and microcalcifications; and separate or combinational CAD systems have been developed for these types of abnormalities. The size, distribution, form, shape and density of these abnormalities are considered as clues in diagnosing their potentially cancerous nature (Tabár & Dean (2005)). Example abnormalities accompanied by their annotations by expert radiologists are shown in Figure 5.

Table 1: Popular publicly available databases in the field of mammography.

database	number of cases	number of images	image format	resolution (bits /pixel)	image mode	view	abnormality	image categories	BIRADS	annotation	origin of database	year
MIAS (Suckling et al. (2015))	161	322	.PGM	8	digitised film images	MLO	all kinds mostly masses	benign, malignant, normal	no	centre and radius of a circle around RoI	UK	2015
DDSM (Heath et al. (2001))	2,620	10,480	.LJPEG	12, 16	digitised film images	MLO, CC	all kinds	benign, cancer, normal, benign without callback(bwc)	yes	contour points of the RoI	USA	1999
BancoWeb LAPIMO (Matheus & Schiabel (2011))	320	1,400	.TIFF	12	digitised film images	MLO, CC	all kinds	benign, malignant, normal	yes	RoI available for a few images	Brazil	2010
INBreast (Moreira et al. (2012))	115	410	.DICOM	14	digital images	MLO, CC	masses, calcifications, distortions, asymmetries	benign, malignant, normal	yes	contour points of the RoI	Portugal	2011
BCDR-F0X (Lopez et al. (2012))	1,010	3,703	.TIFF	8	digitised film images	MLO, CC	all kinds	benign, malignant, normal	yes	lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors	Portugal	2012
BCDR-D0X, BCDR-N01 (Lopez et al. (2012))	724	3,612	.TIFF	14	digital images	MLO, CC	all kinds	benign, malignant, normal	yes	lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors	Portugal	2012
TCGA (BREAST-DIAGNOSIS, TCGA-BRCA) (Clark et al. (2013))	69	88	.DICOM	-	digital images	MLO, CC	all kinds	-	-	-	USA	2001-2009

Table 2: Popular challenges with provided databases in the field of breast histology.

database	number of cases	image format	magnification	slide scanner	resolution (bits/ pixel)	image mode	abnormality	provided assessment	annotation	origin of data	year
ICPR2012 (2017)	5	.bmp	x40	Aperio ScanScope XT	2084×2084	24bit RGB	mitotic nuclei	mitotic locations	centroids of around 300 mitosis and mask in .jpg format	France	2012
				Hamamatsu 2.0HT	2252×2250	24bit RGB					
				Multispectral microscope	2767×2767	gray level					
AMIDA13 (2017)	23	.TIFF, .JPEG	x40	Aperio ScanScope XT	2000×2000	8bit RGB	mitotic nuclei	mitotic locations	centroids of 1157 mitosis and mask in .TIFF format	The Netherlands	2013
MITOS-ATYPIA-14 (2016)	32	.TIFF	x10, x20, x40	Aperio ScanScope XT	1539×1376	RGB	mitosis and nuclear atypia	list of mitosis; list of similar objects to mitosis; nuclear atypia score; mitosis and non-mitosis location; agreement between pathologists	centroids of mitosis and mask in .jpg format; confidence degree in .csv file	France	2014
				Hamamatsu 2.0HT	1663×1485						
CAMELYON16 (2016)	400	multi-resolution pyramid structure	x40, x10, x1	Pannoramic 250 Flash II	pixel size: $0.243\mu m \times 0.243\mu m$	RGB	metastasis	cancerous regions	contours of cancer locations in .xml files and WSI masks	The Netherlands	2016
				Hamamatsu XR C12000	pixel size: $0.226\mu m \times 0.226\mu m$						
TUPAC16 (2016)	500 + axilary datasets	multi-resolution pyramid structure	x40	Aperio XT	highest resolution: 50k×50k	RGB	tumour proliferation	proliferation score; ROC annotation	ROC coordinates with the scores in .csv files	The Netherlands	2016
CAMELYON17 (2017)	200	.TIFF	-	-	-	-	metastasis	micro and macro metastasis; PN stage label; ROC annotation	contours of cancer locations in .xml files and WSI masks	The Netherlands	2016



(a) (b) (c)

Figure 5: (a): mammograms with the annotated mass abnormalities selected from BCDR-F03 database (masses or lumps are the most common symptom of breast cancer); (b): mammograms with the annotated calcification selected from BCDR-F02 database (small deposits of calcium in the breast tissue, called breast calcifications, are common and often associated with benign cases); (c): mammograms with the annotated microcalcifications selected from BCDR-F02 database (shown as much smaller white dots on a mammogram, called clusters of micro-calcifications and are not quite as common and can be a cause of concern). The green boundary represents benign biopsy proven lesions and the red boundary represents malignant biopsy proven lesions.

Radiologists also use a set of intuitive tissue patterns to characterise the appearance of each mammogram manually and estimate breast cancer risk using specific metrics. The reader is referred to (Wolfe (1976); Tabár & Dean (2005); Boyd et al. (2010); D'Orsi (2013); Muhimmah et al. (2006)) for more detailed information about various breast density classification systems. Among these systems, BIRADS [20] has become popular to standardise the mammography report which covers abnormalities and density and more details on the latter are described in Table 3. Each case is assessed to be: *normal* (Assessment

[20]Breast Imaging Reporting and Data System

Table 3: Main breast density categories.

Density percent	BIRADS density class	Tissue appearance
[0%-25%]	I	Predominantly fatty
[25%-50%]	II	Scattered fibro-glandular densities
[50%-75%]	III	Heterogeneously dense
[more than 75%]	IV	Dense

Category (A.C.) = 1), *benign or probably benign* (A.C. = 2 or 3), *suspicious abnormality* (A.C. = 4) or *malignant* (A.C. = 5). The building of systems which can effectively provide automatic detection, segmentation and classification of such lesions based on deep learning methods has become one of the challenging areas in mammographic CAD systems.

3.2. Mass Analysis

The first implementation of deep learning networks in mammographic mass detection/classification was done by Sahiner et al. (1996). The CNN's input images were obtained from manually extracted RoIs [21] by radiologists. With these RoIs, the training data was prepared using two techniques: (i) employing averaging and sub-sampling, (ii) employing texture feature extraction applied to small subregions inside the RoI. They studied the effects of CNN architecture and texture feature parameters on classification of different mammograms and indicated that the input images are more critical than the CNN architecture. However, this was before the use of GPUs and improvement of ANNs and so implementing such CNN was described as computationally intensive and tedious with poor adaptability and limited results. However, later on, inspired by the layer aspect of deep networks along with parallelisable algorithms and properties of GPUs, exploring CNN applications in mammography became more realistic. Petersen et al. (2012) presented a generic multi-scale DAE [22] using a sparsifying

[21] Regions of Interest
[22] Denoising AutoEncoder

activation function for breast density segmentation. They evaluated their results by comparing it to manual BIRADS and Cumulus-like density scoring (Byng et al. (1994)). They showed that multiple scales are effective for learning rich feature representations in the segmentation task. Following Petersen's work, Kallenberg et al. (2016) proposed a CSAE [23] network with sparsity regularisation (both lifetime and population). This architecture expanded the idea of Ranzato et al. (2006) to pixel-wise labelling of large scale images which was able to preserve the spatial layout of the image while avoiding feature overcompleteness. They implied that sparse overcomplete representations are cost-efficient and robust to noise. In a different approach, Jamieson et al. (2012) explored the use of ADNs [24] proposed by Zeiler et al. (2011). ADNs are unsupervised and hierarchical models that use convolution sparse coding and max pooling for image decomposition. They combined the SPM [25] kernel (Lazebnik et al. (2006)) on the inferred feature maps and a linear SVM [26] classifier. They visualised image relationships according to the learned feature information utilising the Elastic Embedding dimension reduction technique. Various depth CNN networks were also tested by Arevalo et al. (2015, 2016). They compared their best obtained results with two baseline descriptors: HOG [27] and HGD [28] and an approach using 17 hand-crafted features. Finally, they reported performance improvement with the combination of both learned and hand-crafted representations. Fonseca et al. (2015) evaluated the performance of the developed HT-L3 CNN, an architecture search procedure technique (Pinto et al. (2009)), on mammograms. The network search space with the proposed options had 729 candidates and it took about 72 hours to screen them in order to find the top 3 performing architectures. By obtaining the best architecture, they performed automatic feature extraction and trained an SVM classifier. Dhungel et al. (2015) presented a

[23] Convolutional Sparse AutoEncoder
[24] Adaptive Deconvolutional Network
[25] Spatial Pyramid Matching
[26] Support Vector Machine
[27] Histogram of Oriented Gradients
[28] Histogram of Gradient Divergence

multi-scale 4-DBN that was combined with a GMM [29] classifier for mass candidate generation. These candidates were fed to a CNN to extract textural and morphological features for the linear SVM classifier (this combination is known as R-CNN). A cascade of two RF [30] classifiers was then applied to the feature set for the inference processes. Performing post processing, regions based on a high overlap ratio were merged as the overall results. Subsequently, Carneiro et al. (2015) fine tuned a CNN pre-trained with ImageNet (Krizhevsky et al. (2012)) using unregistered mammograms and segmented microcalcification and masses. They estimated the patient's risk of developing breast cancer based on BIRADS classification. They concluded that the pre-trained multi-view model is superior to the randomly initialised model in terms of classification since over-fitting of the training data is likely to be caused by a random initialised model. In the recent paper, inspired by their previous work, Dhungel et al. (2016) concluded that the CNN model with pre-training and RF on features from the CNN with pre-training are better than the RF on hand-crafted features and CNN without pre-training.

As a solution to acquiring sufficient data to train a CNN, Sun et al. (2016) hypothesised combining a small amount of labelled data with abundant resources of unlabelled data. The scheme consisted of three modules: i) data weighing (using exponential, Gaussian and Laplacian functions), ii) feature selection (using PCA [31], LDA [32] and MDS [33]) and iii) using their proposed co-training graph based data labelling. With computed weights, the unlabelled data was gradually labelled with a graph based semi-supervised learning method. They implied that their scheme was less sensitive to initial labelled data compared to schemes using the labelled data only, since the additional information for the training was provided by the unlabelled data. Similarly, Kooi et al. (2016) and

[29] Gaussian Mixture Model
[30] Random Forest
[31] Principal Component Analysis
[32] Linear Discriminant Analysis
[33] Multidimensional Scaling

Huynh et al. (2016) took advantage of transfer learning to extract tumour information from medical images via CNNs that were originally pre-trained with non-medical data. Their two-stage classification procedure included detecting candidates for further scrutiny by applying RF and generating likelihood images. These images were then used as seed points for both the reference system and the CNN. They showed that the addition of location, context information and several manually designed features to the network improved the performance. In a similar way, Jiao et al. (2016) proposed a scheme in which a CNN was trained on LSVRC [34] (Deng et al. (2009)) images and fine-tuned on a subset of breast mass images. Then, features of masses were extracted from different hierarchical levels of this model, with the help of which two linear SVM classifiers were trained for the decision procedure. Eventually, in the decision mechanism, the outcomes from different classifiers were fused to complete the classification. Unlike other studies, Samala et al. (2016a) pre-trained CNN on mammography samples to identify specific patterns and transferred this to detect masses in tomosynthesis (an advanced 3D version of mammography). They reported statistically significant performance improvement of deep learning based CADs compared to the feature-based ones.

Classification can be used directly for detection and segmentation. Dubrovina et al. (2016) performed tissue classification with application to the segmentation of pectoral muscle, fibroglandular tissue, nipple and the general breast tissue, which includes fatty tissue and skin. They changed classical fully connected layers in a regular CNN into convolutional layers. In conclusion, they reported significantly faster computation, while preserving the classification accuracy. Fotin et al. (2016) detected soft tissue densities from digital breast tomosynthesis. They compared conventional and deep learning approaches, reporting better CNN performance. Similarly, Kooi et al. (2017) compared a mammography CAD system relying on manually designed features and CNN designed features. They concluded that: i) the CNN based CAD systems out-

[34] Large Scale Visual Recognition Competition

performed the traditional CAD system; ii) there was no significant difference between the model and the radiologists (AUC: 0.85 vs. 0.91); iii) adding manually designed features to the CNN could give very small improvements. In other work, Lévy & Jain (2016) classified pre-segmented masses using different networks from shallow to deep CNNs along with a transfer learning method. They investigated the effect of data augmentation and data context in their work, concluding that double the bounding box of the abnormality is effective in binary classification of masses.

3.3. Microcalcification Analysis

Alongside the CAD models covered already, additional research with regard to microcalcifications, as another major abnormality in mammograms, has been produced. CAD systems are better at detecting and classifying microcalcification than other mammographic abnormalities (Cheng et al. (2003)) as the density of calcium makes detection possible using thresholding. This is not useful for most masses and asymmetries where the density is similar to glandular breast tissue.

The first application of CNN to the detection of microcalcification clusters was performed by Chan et al. (1995). Clusters of micro-calcifications were detected in three main steps: finding SNR-enhanced image by applying enhancement and suppression filters, histogram determination, obtaining signal characteristics and excluding potential signals by thresholding. Subsequently, they trained and investigated the effectiveness of a CNN in detecting and discriminating false signals from true microcalcifications. However, the number of cases they used was limited but they were able to significantly reduce the number of false positive detections. Recently, Wang et al. (2016b), employed a stacked denoising AE to retrospectively analyse microcalcifications with or without masses on mammograms. Microcalcification and mass data were extracted by image segmentation using 41 statistical and textural measurements following the classification. In their work, features were fed into the comparative classifiers rather than the raw images. Its performance and accuracy in clas-

sifying and discriminating breast lesions were compared with SVM, K-nearest neighbour and linear decomposition analysis methods. They reported that the learning power can be enhanced by a combinatorial approach and deep learning based methods are superior to standard methods for the discrimination of microcalcifications. Samala et al. (2016b) used a grid search method to select an optimal CNN architecture for differentiating microcalcification candidates detected during the pre-screening stage. Various filters, filter kernel sizes and gradient computation parameters in the convolution layers were tested to gain the parameter space of 216 combinations. They reported significant improvement on their designed CNN architectures for detection of microcalcifications. Classification of clustered breast microcalcifications into benign and malignant categories was performed by Bekker et al. (2016) which was based on two mammography view-level decisions, allocating separate neural networks for each view. These two view-level soft decisions were then non-linearly combined into a global decision by a single-neuron layer.

3.4. Summary

In summary, introducing deep learning strategies into mammographic analysis has expanded ideas to modify the training process for a wide range of mammographic applications. Detailed information about the implementation of deep learning based methods, covered in this section, is provided in Table 4. Most of these models have tested different network depths and input sizes to address various issues and the majority of models reported improvements over existing state-of-the-art results. An overview of general issues related to deep learning methods in biomedical image analysis is provided by Greenspan et al. (2016); Litjens et al. (2017). For specific case of mammographic analysis, good results are directly related to the correctness of the training data, but the annotations provided by the radiologists are prone to subjectivity. Annotation agreement/disagreement has not yet been included in the currently available datasets which would be helpful for managing errors. In addition, the developed methods are not able to identify the most suitable training exemplars that

contain rich information for a specific task. The developed methods are sensitive to the size of the abnormalities. Nevertheless, to account for morphological variations, abnormalities are first resized to a predefined size to become suitable for the network. Based on the literature review, a combination of deep learning based features and hand-crafted features perform best, but more intelligent combinations are required to be able to respond to the breadth of various mammographic applications.

Table 4: Detailed information for deep learning based methods used in mammography;
*conv: convolution, *fc: fully connected; *FFDM: full field digital mammography; *DBT: digital breast tomosynthesis.

Reference	Task	Data	Number of data	Input size	DL Architecture	Train time	Evaluation
Sahiner et al. (1996)	mass and normal tissue classification	168 mammograms from Department of Radiology, University of Michigan, USA	train: 84 mammograms; test:84 mammograms	16x16 and 32x32	CNN: 2 conv + 1 fc	-	AUC: 0.87
Petersen et al. (2012)	breast density scoring	85 mammograms from a placebo-controlled trial	train: 60,000 patches	28x28	denoising AE: 2 hidden layers with 1,000 neurons each	-	AUC: 0.68
Jamieson et al. (2012)	mass classification	739 image RoIs from University of Chicago Medical Center, USA	-	140x140	4-layer ADN + code book + dictionary histogram image + linear-SVM classifier	-	AUC: 0.71
Fonseca et al. (2015)	density classification	digital images (CC view) from 1,157 subjects at medical centres in Lima, Peru	-	200x200	CNN: 3 conv + SVM classifier	-	AUC: 0.73
Dhungel et al. (2015)	mass classification	410 images from Inbreast	training: 60%; validation: 20%; test: 20% of images	40x40	candidate selection: 4-DBN + GMM classifier; feature learning: cascade of two R-CNNs and two RF classifiers	-	0.96 TPR at 1.2 FPI
		316 images from DDSM-BCRP					0.75 TPR at 4.8FPI
Arevalo et al. (2016)	mass classification	736 film images from BCDR-F03	train: 368; validation: 73; test:295	150x150	2 conv + 1 fc + softmax classifier	about 1.4h (Tesla K40 GPGPU card)	AUC: 0.86
Carneiro et al. (2015)	mass classification	410 images from INbreast	-	264x264	4 conv + 2 fc + softmax classifier	on GPU GeForce GT650M; no extra training samples: 1 hour; 20 additional training samples: 7.5 hours	AUC: 0.91
		680 images from DDSM					AUC: 0.97
Kallenberg et al. (2016)	density scoring	493 mammograms from Dutch breast cancer screening program	train: 48k patches; test: 1,576 cancer/ healthy controls	24x24	3 conv + softmax classifier	-	AUC: 0.59
	texture scoring	668 mammograms from MMHS cohort					AUC: 0.57

Continuation of Table 4

Reference	Task	Data	Number of data	Input size	DL Architecture	Train time	Evaluation
Kooi et al. (2017)	detection of mammographic lesions	nearly 45,000 images from a large scale screening program in The Netherlands	train: 44,090 images; test: 18,182 images	250x250	5 conv + 2 fc + classifier	-	ACC: 0.85
Sun et al. (2016)	mass classification	1,874 pairs in-house full-field digital mammography (FFDM) image database totally 3,158 RoIs	train: 2,400 RoIs; test: 758 RoIs	52x52	3 conv + SVM classifier	-	AUC: 0.88
Kooi et al. (2016)	classification of masses and architectural distortions	397 images from large-scale screening program in The Netherlands	train: 334,752 patches	250x250	5 conv + 2 fc + classifier	-	AUC: 0.87
Huynh et al. (2016)	mass classification	large scale screening program in The Netherlands	train:1,311,272 patches; test: 18,182 patches	250x250	5 conv + 2 fc + classifier	-	AUC: 0.94
Lévy & Jain (2016)	mass classification	1,820 images of 997 patients from DDSM	train: 80%; validation: 10%; test 10% of cases	224x224	Baseline	-	ACC: 0.604
					AlexNet		ACC: 0.89
					GoogleNet		ACC: 0.929
Jiao et al. (2016)	mass classification	600 images from DDSM dataset	-	227x227	5 conv + 2 fc + SVM classifier	-	ACC: 0.967
Samala et al. (2016a)	mass detection	2,282 digitised film and digital mammograms and 324 DBT volumes from University of Michigan and University of South Florida	train: 2,689 mass patches; test: 183 mass patches as true positive	128x128	4 conv + 3 fc	8 days on NVIDIA Tesla K20 GPU	AUC: 0.80
Wang et al. (2016b)	breast lesions classification	-	train: 1,000 images; test: 204 images	-	two layer stacked denoising auto-encoder	-	AUC: 0.87 (on microcalcification features)
							AUC: 0.61 (on mass features)
							AUC: 0.90 (on combinational features)
Chan et al. (1995)	microcalcifications classification	52 mammograms from University of Michigan	train: nearly 1,700 patches; test: nearly 220 patches	16x16	2 hidden layers	-	AUC: 0.9
				20x20			

Continuation of Table 4

Reference	Task	Data	Number of data	Input size	DL Architecture	Train time	Evaluation
Samala et al. (2016b)	microcalcifications classification	64 digital breast tomosynthesis from University of Michigan	train: 4,808 patches; test: 2,220 patches	16x16	2 conv + 2 locally-connected layers + 1 fc	-	AUC: 0.93
Dubrovina et al. (2016)	breast tissue segmentation	40 digital mammograms of mediolateral oblique (MLO) view A leave-one-subject-out cross validation procedure	-	61x61	3 conv + 3 fc	-	Dice coefficient (DC): 0.71
Fotin et al. (2016)	density detection from digital breast tomosynthesis	train: 1864 suspicious mammograms and 339 lesions from DBT	-	256x256	AlexNet	-	ACC: 0.86

4. Deep Learning in Breast Histology Image Processing

4.1. Problem statement

In breast histological imaging, when the biopsied sample is prepared (Veta et al. (2014)), different tissue components are visualised by being stained. The standard staining protocol for breast tissue is H&E which selectively stains nucleic structures blue and cytoplasm pink. After cover-slipping of glass slides, the samples can be digitised with a WSI [35] scanners at a specific magnification. Because of its large size, it is common practice to identify areas of interest in a patch-wise manner to be analysed in CAD systems to decrease computational cost. Figure 6 shows a mammary gland histology slide selected from the University of British Colombia histology repository[36]. This is shown by RoIs at x10, x20 and x40 magnification.

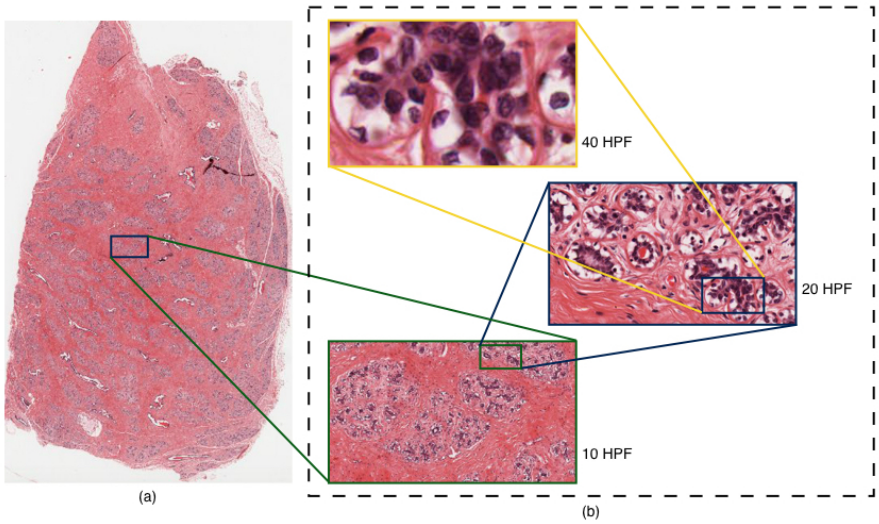
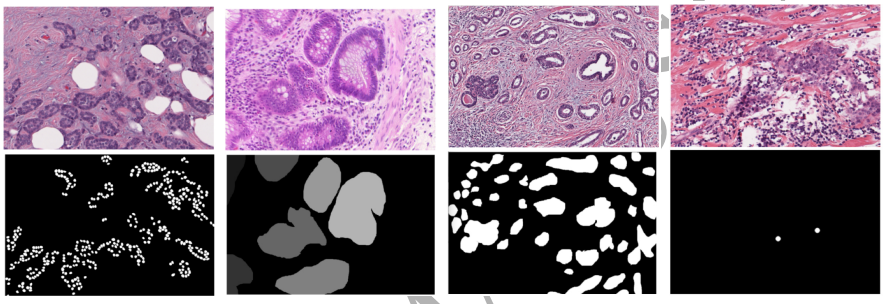


Figure 6: (a): Mammary gland slide scanned with the pixel resolution of $49,440 \times 77,227$; (b): Extracted boxes represent different HPFs from the WSI (x10, x20 and x40 magnification).

[35] Whole Slide Imaging

[36] publicly provided in `http://histo.anat.ubc.ca`

For analysis of breast histopathological images, the Nottingham Grading System (NGS) (Bloom & Richardson (1957)) is recommended by the World Health Organisation. This system is used to predict patient prognosis and provides treatment recommendations. It is derived from the assessment of three morphological features: tubule formation, nuclear pleomorphism and mitotic count (Elston & Ellis (1991)). A numerical scoring system (1-3) is used for the combination of the three grades of tumour differentiation. These features, with the respective annotations [37], are shown in Figure 7. General quantitative



(a) (b) (c) (d)

Figure 7: Top row: different patches extracted from different WSI scans; bottom row: annotations for a specific purpose. It should be noted that inter/intra observer variation in manual annotations can be high. (a) Nuclei detection/segmentation in order to perform pleomorphism grading; (b) Tubule detection/segmentation to assess the degree of structural differentiation in the tissue; (c): Epithelial and Stromal region detection/segmentation that have different significance for prognosis; (d) Mitotic figure detection for grading tumour proliferation i.e. number of mitoses and mitotic activity in tumours correlates with grade and poorer diagnosis.

analysis of breast tissue components on WSI scans includes: nuclei, tubules, epithelium and stroma and mitotic detection. The introduction of deep learning concepts in image processing has provided big datasets along with annotations for specific tasks and some of them are publicly available. Some of these are listed in Table 2. In this review, methods proposed by various deep learning based algorithms for analysing histological components to grade breast cancer

[37] publicly provided at `http://www.andrewjanowczyk.com`

on histology data are covered.

4.2. Nuclei analysis

Breast epithelial nuclei usually look different in shape, size, texture and mitotic count according to nuclei life cycle and malignancy level of the disease. Nucleic pleomorphism has important diagnostic value for predicting the existence of disease and its severity. Inspired by Cireşan et al. (2013), Xu et al. (2014) developed an unsupervised two-layer SSAE [38] framework for nuclei classification. An SAE was trained to capture primary feature activations on raw input patches. Then, these primary features were fed to another SAE to learn secondary features for each of the primary features. Subsequently, being analysed by a classifier, the secondary features were mapped to the respective labels. They compared "SSAE + softmax", "PCA + softmax" and "a single layer SAE + softmax" frameworks for the task of patch-wise classification. Their results showed the "SSAE + softmax" out-performed the other methods on their own dataset. They extended this framework to automatically detect multiple nuclei by computing locally maximal confidence scores across the entire image (Xu et al. (2016b)). They further compared their model with several nuclei detection methods and concluded that this framework can provide accurate seed points for developing cell-by-cell graph features. Characterising cellular topology features on tumour histology was reported to be a promising advantage of this framework. Janowczyk & Madabhushi (2016) performed a comprehensive study of deep learning approaches for 5 different breast tissue tasks in histology image processing. They provided additional online material and implementations[39] and tried to decrease computational cost caused by interrogating all the image pixels (Janowczyk et al. (2016)). To this end, a resolution adaptive deep hierarchical learning scheme was suggested in which higher levels of magnification were used when needed. As a result, they were able to reduce the

[38] Stacked Sparse AutoEncoder

[39] http://www.andrewjanowczyk.com

computation time by about 85% on ER+ [40] breast cancer images. Xing et al. (2016) performed nucleus segmentation while preserving the shape by generating probability maps using CNN models and applying selection-based sparse shape and local repulsive deformable models. They showed that this approach is applicable to different H&E stained histopathology images, evaluating on three histopathology image datasets from different tissues (including breast tissue) and stain preparations. Veta et al. (2016b) computed statistics of individual nuclei and surrounding regions by training a deep-CNN [41] model on tumour region images with known nuclei locations. They were able to do so directly from the image data without the need for nuclei segmentation. Xie et al. (2015) proposed a modified CNN model for cell detection by using a structured regression layer instead of a classifier. This way, they aimed to encode topological information which was ignored in the conventional CNN because of the coherency in labelled regions. Handling inhomogeneous background noise and size and shape variations were the significant strength of their method.

4.3. Tubules analysis

Identifying tubule nuclei from WSIs in order to calculate the ratio of tubule nuclei to the overall number of nuclei (tubule formation indicator) was studied by Romo-Bucheli et al. (2016). They used a customised CNN to quantify tubule score in ER+ breast cancer WSIs. Patches of nuclei candidates, that were extracted by the customised CNN, were manually labelled as containing a tubule or not. Subsequently, a deep learning based network was trained to detect and classify tubule nuclei. They concluded that the tubule formation indicator correlated with the likelihood of cancer recurrence.

[40] Estrogen-Receptor-Positive
[41] DCNN

4.4. Epithelial and Stromal region analysis

For this task, Xu et al. (2016a) presented a patch based DCNN [42] approach for distinguishing epithelial and stromal components in H&E stained tissue images. The images were over-segmented into small regions using two different superpixel algorithms (the Ncut [43] algorithm and the SLIC [44] algorithm). Evaluating the comparative strategies, the combination of DCNN with the Ncut-based algorithm and a SVM classifier led to the best results. Bejnordi et al. (2017) trained two deep CNNs inspired by VGGNet. The only modification was that they replaced the two fully connected layers with convolutions to allow arbitrary input sizes to be fed to the network. In their work, the first CNN model was trained to classify the WSI into epithelium, stroma, and fat. The second CNN model was trained on the resulting stromal areas to classify the stromal regions as normal or cancerous.

4.5. Mitotic activity analysis

To quantify the locality and proliferative activity of breast tumours, mitotic count is estimated as the number of mitoses in an area of $2mm^2$ (usually using microscope magnification of $\times 40$) and reported as the MAI [45] (Van Diest et al. (2004)). This gives an evaluation of the aggressiveness of the tumour. Mitosis detection is challenging due to the small size with a large variety of shape configurations of mitoses. In H&E stained breast cancer sections, mitoses are hyperchromatic objects lacking a clear nuclear membrane with their own specific shape properties. Inspired by the outstanding results for using patch-driven CNN in image classification and segmentation (Cireşan et al. (2012a,b)), Cireşan et al. (2013) used the deep max-pooling CNN architecture operated directly on raw RGB pixels. They tried to reduce the deep neural network's variance and bias by averaging the outputs of multiple classifiers with different architectures

[42] Deep CNN
[43] Normalised Cut
[44] Simple Linear Iterative Clustering
[45] Mitotic Activity Index

along with using rotational invariance. Their method won the ICPR12 competition with the highest F-score and precision. With the same approach plus employing a Multi-column CNN, the same team won the AMIDA13 competition in which three CNNs were trained on nearly 20 million samples (Veta et al. (2015)). The output probabilities of the CNNs were averaged and used as the final result. Wang et al. (2014a,b) fused a lightweight CNN with hand-crafted features (morphological, statistical and textural sets) for each candidate region defined by thresholding. Extracting these features independently, a cascade of two random forest classifiers was combined and trained. They showed that the integrated approach resulted in superior detection accuracy compared to individual deep learning or hand-crafted feature based approaches. In the same way, Malon & Cosatto (2013) combined manually segmentation-based nuclear features (colour, texture, and shape) with the features extracted by a LeNet-5 architecture (LeCun et al. (2010)). Reported advantages were: handling the appearance varieties in mitotic figures, decreasing sensitivity to the manually crafted features and thresholds. Chen et al. (2016a) suggested a deep cascade neural network with two phases. In the first phase, a 3-layer CNN was utilised to retrieve probable mitosis candidates and in the second phase, three CaffeNet-based CNNs (Jia et al. (2014)) were used to detect mitotic cells in all positive samples determined by the first CNN. In other work, Chen et al. (2016b) implemented a deep regression network along with transferred knowledge for this task and showed the efficiency of their proposed approach in dealing with automatic mitosis detection.

To overcome the bottleneck of access to a large number of annotated training samples for mitosis detection with deep CNNs which is more critical compared to the other tasks, Albarqouni et al. (2016) presented a new concept for learning from crowds and generating ground-truth labelling from non-expert crowd sourced annotations. In their proposed data aggregation framework, they trained a multi-scale CNN model using gold-standard annotations. Then, in the second step, using the incoming unlabelled image, aggregation schemes were integrated into CNN layers via an additional crowdsourcing layer (AggNet).

AggNet could produce a response map, refine the CNN model by filtering out weak responses and simultaneously generate a ground-truth by majority crowd sourced votes. They analysed the behaviour of CNN with and without aggregation and confirmed that aggregation and deep learning from crowd annotations was robust to noisy labels (multiple different labels for the same sample). They claimed that not only could deep CNNs be trained with data collected from crowdsourcing, but also it positively influenced the CNN performance. Such results could be valuable in giving insight into the functionality of deep CNN learning from crowd sourced annotations. Veta et al. (2016a) presented an analysis of the object-level inter-observer agreement on mitosis counting. They compared the performance of their deep learning based mitosis detection which was trained on the AMIDA13 database with the performance of expert observers on an external dataset. They described disagreement among pathologists which in some cases was significant. They concluded that automatic mitosis detection performed in an unbiased way and provided substantial agreement with human experts.

4.6. Other tasks in breast digital histopathology image processing

Detection of invasive ductal carcinoma [46] in WSI for the estimation of tumour grading and the prediction of patient outcome was done by Cruz-Roa et al. (2014). Using a three-layer CNN, they evaluated their network over a WSI dataset from 162 patients diagnosed with IDC. Comparing their results with the outcome from hand-crafted image features (colour, texture and edges, nuclear textural and architecture) with a random forest classifier, they reported their best quantitative results for automatic detection of IDC regions in WSI.

Wang et al. (2016a) investigated the applicability of various CNNs (AlexNet, GoogLeNet, VGG16 and FaceNet) in breast cancer metastases detection in resected sentinel lymph nodes (first lymph node to which cancer cells are most likely to spread to). They won the Camelyon16 competition for WSIs clas-

[46] IDC

sification and tumour localisation. In their results, the two deeper networks (GoogLeNet and VGG16) achieved the best patch-based classification performance with x40 magnification. Their results also demonstrated that the combination of deep learning methods with pathologist's interpretation could reduce the error rate by 85% which is a significant improvement in diagnostic accuracy. Similarly, Litjens et al. (2016) identified slides that did not contain micro/macro-metastases. Accordingly, a CNN was trained to obtain per-pixel cancer likelihood maps and segmentations in whole-slide images rather than a patch-by-patch classification. Janowczyk et al. (2017) attempted to evaluate Stain Normalisation via Sparse AutoEncoders under different circumstances: i) in different concentrations of H&E in the same tissue section; ii) with the same slides being scanned multiple times on different platforms. In addition, they compared their proposed approach with other colour normalisation methods and reported outperforming the alternative approaches. Their approach standardised colour distributions of a test image to a single template image and increased robustness to different sources of variance like specimen thickness, stain concentration and scanner.

4.7. Summary

Deep learning algorithms try to emulate the way histopathologists examine whole tissue slides. Several studies have compared the performance of deep learning methods to the performance and interobserver agreement of expert pathologists (Giusti et al. (2014)). Histopathologists analyse the image at low magnifications and then perform more sophisticated analysis on some specific areas requiring more detailed information under higher magnification. Selecting appropriate magnifications in deep learning methods remains a challenge. The identification of the best training set containing richly informative exemplars is another concern. However, the lack of readily available annotated data for digital histopathology analysis is not as critical as for mammography since one WSI typically contains trillions of pixels from which hundreds of targeted examples can be extracted. Moreover, some competition challenges (see Subsection 1.3)

have provided access to publicly available data which are systematically annotated. From the literature, it can be concluded that, deep learning approaches have proven capability in discriminating between the targeted classes by combining both feature discovery and implementation. The strategy of combining both deep learning based and hand-crafted features has enabled the possibility of achieving state-of-the-art performance when using AI for the interpretation of x-ray and histology images of breast cancer. Although these deep learning based approaches have demonstrated promising results, there is still progress to be made to reach clinically acceptable results. Detailed information about the implementation of deep learning based methods, covered in this section, is provided in Table 5.

Table 5: Detailed information for deep learning based methods used in breast histopathology;
*conv: convolution, *fc: fully connected; *FFDM: full field digital mammography; *HPF: high power field.

Reference	Task	Data	Number of data	Input size	DL Architecture	Train time	Evaluation
Xu et al. (2014)	nuclei classification	17 patient cases containing 37 H&E images at Case Western Reserve University	train: 14421 nuclei and 28032 non-nuclei patches; test: 2000 nuclei and 2000 non-nuclei patches	34x34	SSAE with 2 hidden layers (500 and 100 neurons respectively) + classifier	-	F-score: 0.82
Xu et al. (2016b)	nuclei detection	537 H&E images corresponding to 49 lymph node-negative and estrogen receptor-positive breast cancer (LN-, ER+BC) patients at Case Western Reserve University	train: 37 images; test: 500 images	34x34	SSAE with 2 hidden layers (400 and 225 neurons respectively) + classifier	2.15 hours	F-score: 0.8449
Xing et al. (2016)	nucleus segmentation	anonymous	train: 35 images; test: 35 images	45x45	CNN: 2 conv + 3 fc + classifier	-	F-score: 0.78
Xie et al. (2015)	cell detection	32 images from The Cancer Genome Atlas (TCGA) dataset	train: 16 images; test: 16 images	49x49	CNN: 2 conv + 3 fc	-	F-score: 0.913
Janowczyk et al. (2016)	nuclear segmentation	anonymous	141 ER+ breast cancer images	32x32	AlexNet	-	F-score: 0.84
Veta et al. (2016b)	computing nuclear area statistics	39 slides from patients with invasive breast cancer from University Medical Center Utrecht, The Netherlands	train: 14 cases; validation: 7 cases; test: 18 cases	96x96	CNN: 8 conv + 2 fc + classifier	-	coefficient of determination: 0.77
Xu et al. (2016a)	Epithelial-Stromal segmentation	106 H&E images from Netherlands Cancer Institute (NKI)	train: 69 images; test: 37 images	50x50	2 conv + 2 fc + Softmax classifier	-	F-score: 0.8521
		51 H&E images from Vancouver General Hospital (VGH)	train: 36 images; test: 15 images				F-score: 0.891
Romo-Bucheli et al. (2016)	tubule detection and classification	174 ER+ breast cancer images	train: 163 patient WSI; test: 11 patient WSI	64x64	CNN: 3 conv + 3 fc + classifier	-	F-score: 0.59

Continuation of Table 5

Reference	Task	Data	Number of data	Input size	DL Architecture	Training time	Evaluation
Bejnordi et al. (2017)	classification of tissue into epithelium, stroma, and fat.	646 H&E sections (444 cases) in the Breast Cancer Radiology Evaluation and Study of Tissues (BREAST) Stamp Project	training: 270 WSIs; validation: 80 WSIs; test: 296 WSIs	224x224	CNN1: VGG-net with 11 layers	-	ACC: 0.95
	stromal regions classification				CNN2: VGG-net with 16 layers		ACC: 0.921
	breast cancer classification				CNN1 + CNN2 + random forest classifier		ROC: 0.92
Cireşan et al. (2013)	mitosis detection	ICPR12 mitosis dataset	train: 35 HPFs; test: 15 HPFs	101x101	DNN1: 5 conv + 2 fc + softmax classifier; DNN2: 4 conv + 2 fc + softmax classifier	-	F-score: 0.782
Malon & Cosatto (2013)	mitosis detection	ICPR12 dataset	train: 35 HPFs; test: 15 HPFs	72x72	CNN (2 conv + 2 fc + SVM classifier) + hand-crafted features	-	F-score (on colour scanners) = 0.659 F-score (on multispectral scanner) = 0.589
Wang et al. (2014a)	mitosis detection	ICPR12 dataset	train: 35 HPFs; test: 15 HPFs	80x80	cascade of CNN (2 conv + 1 fc + RF classifier) and hand-crafted features	about 18 hours with GPU	F-score: 0.7345
		AMIDA13 dataset	train: 12 HPFs; test: 11 HPFs			-	F-score: 0.319
Albarqouni et al. (2016)	mitosis detection	AMIDA13 dataset	train: 311 HPFs; validate: 60 HPFs; test: 295 HPFs	33x33	3 conv + 1 fc	-	AUC: 0.8695
Chen et al. (2016b)	mitosis detection	ICPR12 mitosis dataset	train: 35 HPFs; test: 15 HPFs	480x480	CNN: 5 conv + 3 fc + classifier	-	F-score: 0.79
Wang et al. (2016a)	breast cancer metastasis detection and localisation	Camelyon16	train: 270 WSI; test: 130 WSI	256x256	GoogLeNet	-	ACC: 0.984
					AlexNet		ACC: 0.921
					VGG16		ACC: 0.979
					FaceNet		ACC: 0.968

Continuation of Table 5

Reference	Task	Data	Number of data	Input size	DL Architecture	Training time	Evaluation
Litjens et al. (2016)	breast cancer metastasis detection in sentinel lymph nodes	digitised H&E-stained slides from 271 patients at 3D Histech, Budapest, Hungary	train: 98 slides; validation: 33 slides; test: 42 slides	128x128	4 conv + 2 fc	per epoch: 200minutes using GeForce GTX970	AUC: 0.88
Cruz-Roa et al. (2014)	invasive ductal carcinoma (IDC) detection	169 cases from the Hospital of the University of Pennsylvania and The Cancer Institute of New Jersey	train: 82,883 patches; validation: 31,352 patches; test: 50,963 patches	100x100	2 conv + 2 fc + logsoftmax classifier	-	F-score: 0.718
Janowczyk et al. (2017)	stain normalization	anonymous	train: 200 images; test: 25 images	32x32	AE: 2 layer, first layer with 1,000 hidden neurons, second with 10 hidden neurons	5 hours using a Nvidia M2090 GPU with 512 cores at 1.3 GHz	error: 0.047
Janowczyk & Madabhushi (2016)	nuclei segmentation	anonymous	train:100; test:28	32x32	AlexNet	on Tesla M2090 GPU + CUDA.5, without cuDNN: 22 hours; on Tesla K20c + CUDA.7 with cuDNN: 4 hours	F-score: 0.83
	epithelium segmentation		train:34; test: 8				F-score: 0.84
	tubule segmentation		train:21; test:5				F-score: 0.83
	mitosis detection		-				F-score: 0.53
	invasive ductal carcinoma detection		-				F-score: 0.76

5. Biological Mammography Histology Association

From a biological point of view, it has been long recognised that in the breast the underlying differences in cellular architecture and nuclear morphological alterations lead to tissue changes and the formation of masses, microcalcifications or other abnormalities (Boyd et al. (1992)). Tumour morphology in histology images can reflect some of all possible molecular pathways occurring in tumour cells. In other words, these biological pathways and cellular alterations contribute to the structural and functional attributes in radiographic images (Madabhushi & Lee (2016)), which is represented by both mammography and histology.

There are a number of publications which have provided evidence for the association between radiological and histological risk factors (Ghosh et al. (2012); Pang et al. (2015); Holland & Hendriks (1994); Britt et al. (2014); Lamb et al. (2000); Beck et al. (2011); Sun et al. (2014); Dos Santos et al. (2016)). Britt et al. (2014) defined the association between histopathological characteristics and mammographic density based on the changes in epithelial cells, stromal cells, the extracellular matrix, immune infiltrating and the roles of each cell type in breast cancer initiation and progression. In a case study, Holland & Hendriks (1994) investigated the link between mammographic and histologic appearances in different types of DCIS [47]. They found that linear, branching, granular and coarse microcalcifications corresponding to the amorphous type calcifications in histology were associated with high grade DCIS. While multiple clusters of fine granular microcalcifications corresponding to the clusters of laminated, crystalline calcifications in histology were associated with well-differentiated DCIS. Lamb et al. (2000) reported that larger tumour sizes on mammography resulted in higher grades in histology. However, spiculated margins on a mammogram, associated with acoustic shadowing on ultrasound, were documented as low-grade tumours while most high-grade tumours had a poorly

[47] Ductal Carcinoma In-Situ

defined margin. Malignant-type microcalcifications were mostly seen in mammograms associated with high-grade tumours. Ghosh et al. (2012) also reported that dense areas of the breast in mammograms are different from non-dense areas from a histological point of view so that investigation of both epithelial and stromal components were important in understanding the association between mammographic density and breast cancer risk. Identification of histologic image features that can be predictive of breast cancer survival were studied by Beck et al. (2011). Sun et al. (2014) investigated the relationship between breast tissue composition and age, body mass index, and tumour grade. They concluded that morphological features of breast tissue could influence breast cancer etiology. Dos Santos et al. (2016) investigated biological aspects of immunohistochemical and histological composition of dense and non-dense breast tissue in 18 women. Based on their reported findings, the number of TDLU [48] was higher in dense tissue. They concluded that both stroma fibrosis and epithelial proliferation were responsible for higher mammographic density, so that no proliferative lesions with atypia were found in non-dense tissue, while epithelial atypia was observed in some dense areas. In addition, proliferative lesions without atypia and non-proliferative lesions were found in both tissues, but more frequently in dense tissue. Extensive or moderate fibrosis in dense tissue was the other differentiation with non-dense tissue histological characterisation.

Tot & Tabár (2011) investigated correlation of the radiologic and histopathologic findings. They assessed the clinical relevance of several parameters, that are often verified by pathologists and documented in large-format histological sections, such as: size of the cancer, the extent of the disease, the distribution of lesions and tumour heterogeneity. They concluded that a comprehensive radiological-pathological correlation was the most informative way of early breast cancer diagnosis so that diagnostic failure was due to insufficient radiological-pathological correlation.

Despite biological interpretations, the internal information, generated in

[48] Terminal Ductal Lobular Units

deep networks used in CAD systems, has the potential to add to our knowledge about the existing association between histological compositions and mammographic phenotypes. This is discussed in more details in the next section.

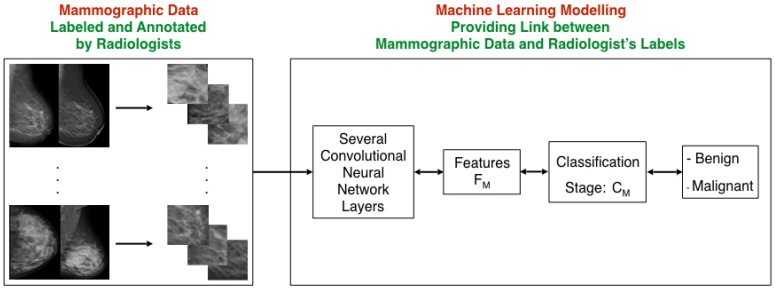
6. Conclusions and Future Trends

6.1. Conclusions

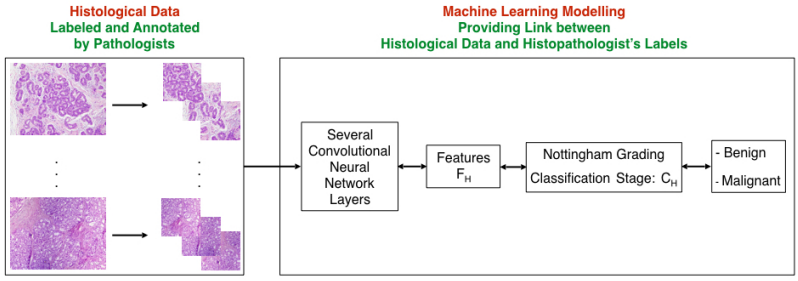
As explained in Sections 3 and 4, information for estimating breast cancer stage and risk can be obtained using different imaging modalities. Methods focused on in this review include histological appearance of the breast nuclei and epithelium detected in biopsy specimens, radiological appearance of abnormality and parenchymal patterns in densities revealed by mammograms. These imaging modalities that manifest across multiple different length scales (micro and macro imaging scales) offer a wide range of information and clinicians combine these heterogeneous sources of data for better disease diagnosis and treatment planning. However, as described in Section 2, many cases with suspicious abnormal findings in mammography who went for further biopsy, eventually were found to have unnecessary biopsies. Motivated by the biological association between mammography and histology (covered in Section 5) and considering the capabilities of deep learning based models in learning from raw data suggests a methodology to potentially reduce biopsies. It is assumed that the appearance of mammographic abnormalities can be linked to specific histological information and can predict how the micro-biological changes are reflected in macro-images.

6.2. Mammography-Histology-Phenotype-Linking-Model

Finding radiological-histopathological correlation/association has been investigated from a biological point of view as described in Section 5. Most of these epidemiological studies are based on empirical observations and statistical risk analysis. However, to the best of our knowledge, a computer based model of such correlation/association is not yet developed. In this paper, we have tried to cover this research question and propose a general framework for fully automatic linking of mammographic and histologic phenotypes.



(a) $Model_M$



(b) $Model_H$

Figure 8: Separate CAD models for (a) mammography ($Model_M$) and (b) breast histology ($Model_H$).

Figure 8, shows the development of automatic CAD systems for the mammography and histology data analysis (covered in Sections 3 and 4, respectively), which are expected to use modern machine learning techniques (e.g. deep learning, convolutional neural networks, autoencoders, etc.) to determine a set of mammographic (F_M) and histological (F_H) phenotypes/features/abstracts, which are discriminative in various image processing tasks such as detection, segmentation and classification. It should be noted that the modelling will be an optimisation process and for the training data the labels are used to estimate the model parameters and generate appropriate features.

Once the mammographic and histological models are estimated, they can be used to generate patient matched mammographic and histological feature/phenotype weighting and their relationship can be estimated by developing a model link-

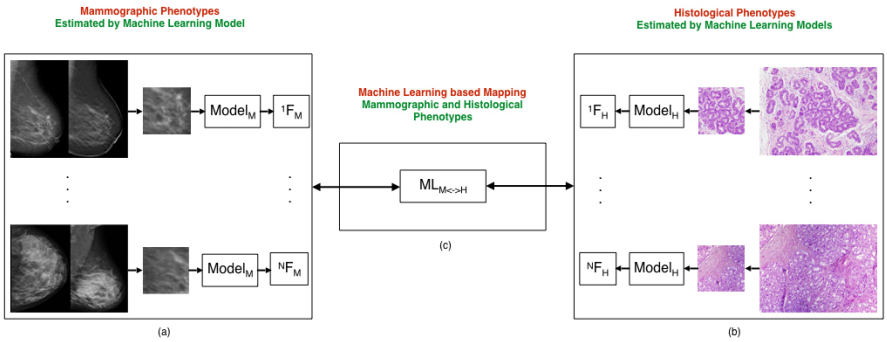


Figure 9: General framework for developing the Mammography-Histology-Phenotype-Linking-Model. (a) $Model_M$: mammographic machine learning based model creating mammographic features (F_M); (b) $Model_H$: histological machine learning based model creating histological features (F_H); (c) the $ML_{M<->H}$ model for providing associations between mammographic and histologic features.

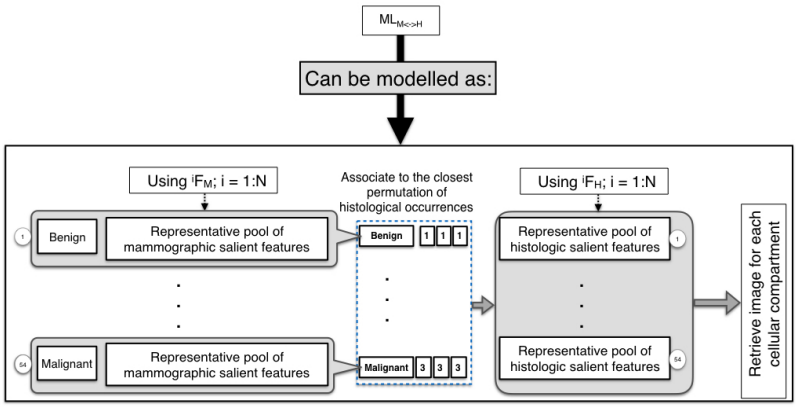


Figure 10: Proposed methodology of developing the Mammography-Histology-Phenotype-Linking-Model using deep learning based approaches. $Model_M$: mammographic deep learning based model, $Model_H$: histological deep learning based model, F_M : mammographic high level deep learning based features, F_H : histological high level deep learning based features, $ML_{M<->H}$: relationships between the mammographic and histologic phenotypes. This can be achieved by: (1) creating different clusters based on permutation of 3 histological score occurrences; (2) associating created pools of deep learning based features to the proper cluster based on the available annotations and making discriminative clusters; (3) matching representative pools of mammographic and histologic features; (4) by using high level histologic abstracts and performing deconvolution/decoding of $Model_H$, morphological approximations can be estimated.

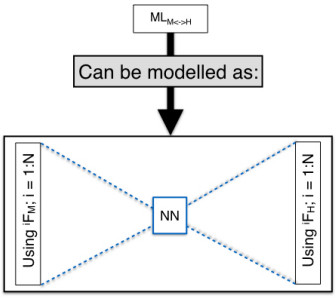


Figure 11: Proposed methodology of developing the Mammography-Histology-Phenotype-Linking-Model using deep learning based approaches. $Model_M$: mammographic deep learning based model, $Model_H$: histological deep learning based model, F_M : mammographic high level deep learning based features, F_H : histological high level deep learning based features, $ML_{M<->H}$: relationships between the mammographic and histologic phenotypes. This can be achieved by using matched F_M and F_H features as input and output of a Neural Network (e.g. an autoencoder)

ing the two based on machine learning techniques (see Figure 9). The automatic extraction of morphological/appearance features from mammographic and histological images and building a map between these based on a large dataset form essential parts in developing such a model. One possible solution for developing such a *"Mammography-Histology-Phenotype-Linking-Model"* or in short *"$ML_{M<->H}$"* is shown in Figure 10. In this approach, a mammographic model ($Model_M$) can be trained, which is based on minimising the difference between NHS/BIRADS labels provided by expert radiologists and those predicted/estimated by the model. Subsequently, salient deep and high level features are generated (F_M) and a pool of deep learning based features is created for each individual image. Similarly, a histological model ($Model_H$), which is based on minimising the difference between NHSBSP histopathology reports and those predicted/estimated by the model can be trained. Using the Nottingham Grading System (NGS), this model is able to predict scores of 1-3 for three cellular components important in breast histology diagnosis (i.e. nuclei, tubules and mitoses). At the same time, this model is capable of creating a pool of high level and deep learning based features for each component. Permu-

tation of the 3 scores for each histological component with 2 possible outcomes (benign and malignant) will result in 54 possible occurrences ($3^3 \times 2$). Therefore, 54 clusters can be formed, although it should be noted that some might only be sparsely populated. To develop the $ML_{M<->H}$ model, the starting point is to generate a set of matched mammographic and histologic features/abstracts created by $Model_M$ and $Model_H$, respectively. To achieve this, the created mammographic features are associated with their respective NGS cluster and a pool of representative mammographic features for a specific cluster is formed. Meanwhile, each cluster in the permuted set contains a pool of previously generated histologic features. A mapping between the two feature spaces will be provided considering that mammographic and histologic data are provided for individual cases. Eventually, machine learning techniques are exploited to retrieve different morphological appearances for each cluster, resulting in the final $ML_{M<->H}$ model.

An alternative approach to develop the $ML_{M<->H}$ model (see Figure 11) avoids the need for clustering and basically uses the matched F_M and F_H features, as respectively input and output to build a simple autoencoder model which maps the two domains through a reduced set of features. The downside of such a model is the lack of clinical reference of the reduced feature set, whilst the advantage is the simplicity of the resulting $ML_{M<->H}$ model.

The final stage of development is to use unseen mammographic cases to predict the histological classification based on the Nottingham Grading Scheme. An overall predictive model is shown in Figure 12. An unseen mammographic case can be processed in a number of ways, which all require initial processing towards a mammographic phenotype/feature (F_M) representation. The mammographic classification stage (C_M) leads to mammographic NHS/BIRADS classification. Using appropriate similarity measures in the $ML_{M<->H}$ model, predicted feature sets are associated to the closest cluster which results in NHSBSP Histopathology Reporting Form classification (or the NN model) and the set of matched abstract features (F_H), which with $Model_H$ leads to the estimation of histological appearance/ phenotypes.

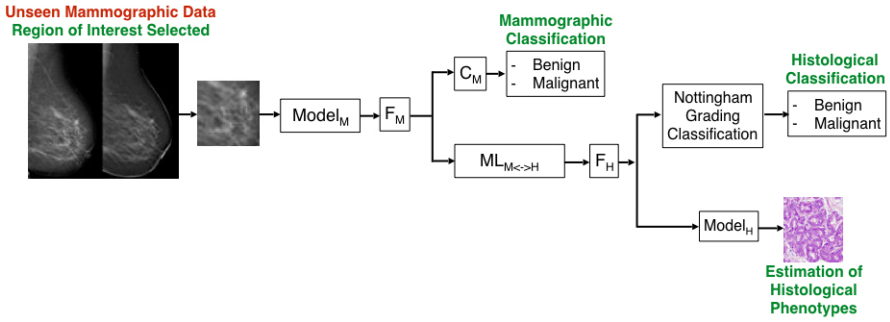


Figure 12: Using the Mammography-Histology-Phenotype-Linking-Model ($ML_{M<->H}$) for unseen cases. $Model_M$: Mammographic machine learning based model, $Model_H$: histological machine learning based model, F_M : mammographic phenotypes, F_H : histological phenotypes, C_M : mammographic classification, $ML_{M<->H}$: relationships between mammographic and histologic phenotypes.

6.3. Possible challenges

Despite the promising results obtained by deep learning approaches, there are remaining challenges for the development of the $ML_{M<->H}$ model, which include:

1. *Data availability*: The first and most basic challenge is the availability of a large number of training samples specifically for this application since the mammograms and histological images should be matched for individual women. The number of samples should be large enough for deep network training purposes. However, existing datasets can be used in the pre-training stage to compensate for the lack of annotated mammography/histopathology data. Appropriate data might be available on existing PACS (Picture Archiving and Communication Systems). As explained in Section 1.1, women are sent for mammography imaging prior to biopsy. Therefore, for the existing histological data, the respective mammograms and the corresponding diagnostic reports exist in digital structured archives, but ethical and research governance agreement and approval will be necessary.
2. *Combinational ground truth*: Appropriate ground truth for the validation

865 part of each individual, mammography and histology image processing, task should be defined systematically. For example the annotations required for breast tissue segmentation in mammograms (characterisation) is different from the annotation required for cancer and non-cancer classification of the tissues. The annotation required for the mitotic count 870 (characterisation) in histopathology is very different from the annotation required for cancer and non-cancer classification of the regions (classification). For associating abnormal phenotypes in a mammogram to characteristics of the tumour in histology, some specific annotations (location and type of abnormality along with locations of nuclei, mitosis count and 875 tubules morphology) are of interest.

3. *Subjectivity of annotations*: If possible, annotations should be provided by different radiologists and histopathologists to accommodate subjective variations. This inter/intra expert variation then needs to be taken into account (Irshad et al. (2014)).

880 4. *Robustness to data acquisition methods*: The issue of robustness to various clinical/technical conditions should be addressed so that gradually more datasets can be added. These variations include: different scanners used for image acquisition; different lighting conditions; various size and views in both mammography and histology; different staining appearance char- 885 acteristics and magnification factors in histology. The developed method should be robust with respect to such variabilities and appropriate normalisation techniques could facilitate this.

5. *Interpretability of model layer information*: Unlike hand-crafted features that provide transparent information, which are more intuitive and inter- 890 pretable to clinicians and researchers, deep learning driven features rely on filter responses solicited from a large amount of training data which suffer from a lack of direct human interpretability. Therefore, approaches to blend domain inspired features with deep learning based features can be taken into consideration in order to take advantage of domain knowledge 895 while enabling the classifier to discover additional features.

6. *Association making algorithms*: New algorithms for combining mammographic and histologic measurements should be designed, which is a more detailed version of the high level descriptions provided in Section 6.2. By finding and visualising a logical association between outcome features introduced by deep networks and the salient diagnostic features incorporated in conventional machine learning based CAD systems, a subset of clinically salient features can be determined. Such association making algorithms, as the novel part of mammographic-histologic linking map introduced in this paper, is an open challenge for future research. One alternative approach to tackle this challenge is by combining image data with text reports as addressed by Shin et al. (2015) while expanding this to the field of radiology and histology in order to mine the semantic interactions between radiology and histology images and the corresponding reports.
7. *Clinical feedback*: More evidence and feedback regarding the results of clinical applications using the developed models will need to be provided by clinicians. Close cooperation between radiologists, pathologists and computer scientists will be necessary for the optimum management of data, analysing the performance of developed methods in a clinical setting with feedback from the radiologists and histopathologists throughout the research process.

6.4. Clinical relevance

The described linking map is expected to reduce the need for further biopsy when the mammographic abnormality is deemed benign as it is reported from a biological point of view (Tot & Tabár (2011)). This association map could contribute to clinical decision making, diagnosis and treatment management. This may also improve the capabilities of computer aided prognosis systems to find patients susceptible to specific breast cancer types at an early stage and as such decrease time before diagnosis, expense and stress. This exploratory research work could be further extended to finding the link between mammography phe-

notypes, histological signatures and protein/gene expression and so be useful for predicting recurrence of and survival after breast cancer. Other imaging modalities for breast imaging, such as MRI and Ultrasound could be exploited in the development of a linking map. This could also cover various ethnic populations and links to breast cancer pathways. It could identify sub-cellular patterns of involved proteins and their locations for cancerous and non-cancerous tissues by avoiding the need for invasive biopsy sampling. Identification of the factors responsible for high-risk histological changes can potentially lead to modelling of disease appearance, better prediction of disease aggressiveness and finally patient outcome.

References

References

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, *35*, 1313–1321.

American-Cancer-Society (2016). What are the key statistics about breast cancer?

AMIDA13 (2017). Assessment of Mitosis Detection Algorithms. MICCAI Grand Challenge.

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2015). Convolutional neural networks for mammography mass lesion classification. In *IEEE 37th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)* (pp. 797–800).

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *127*, 248–257.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, *3*, 108ra113–108ra113.

Bejnordi, B. E., Linz, J., Glass, B., Mullooly, M., Gierach, G. L., Sherman, M. E., Karssemeijer, N., van der Laak, J., & Beck, A. H. (2017). Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In *arXiv preprint arXiv:1702.05803*.

Bekker, A. J., Greenspan, H., & Goldberger, J. (2016). A multi-view deep learning architecture for classification of breast microcalcifications. In *13th International Symposium on Biomedical Imaging (ISBI)* (pp. 726–730). IEEE.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*, 1–127.

Bloom, H., & Richardson, W. (1957). Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, *11*, 359–377.

Boyd, N., Jensen, H. M., Cooke, G., & Han, H. L. (1992). Relationship between mammographic and histological risk factors for breast cancer. *Journal of the National Cancer Institute.*, *84*, 1170–1179.

Boyd, N. F., Martin, L. J., Bronskill, M., Yaffe, M. J., Duric, N., & Minkin, S. (2010). Breast tissue composition and susceptibility to breast cancer. *Journal of the National Cancer Institute.*, *102*, 1224–1237.

Boyer, B., Balleyguier, C., Granat, O., & Pharaboz, C. (2009). CAD in questions/answers: Review of the literature. *European Journal of Radiology*, *69*, 24–33.

Breast-Cancer-Biopsy (2016). Breast Cancer Biopsy.

Britt, K., Ingman, W., Huo, C., Chew, G., & Thompson, E. (2014). The pathobiology of mammographic density. *Journal of Cancer Biology and Research*, *2*, 1021.

Byng, J. W., Boyd, N., Fishell, E., Jong, R., & Yaffe, M. J. (1994). The quantitative analysis of mammographic densities. *Physics in Medicine and Biology*, *39*, 1629.

CAMELYON16 (2016). ISBI challenge on cancer metastasis detection in lymph node.

CAMELYON17 (2017). automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections.

Carneiro, G., Nascimento, J., & Bradley, A. P. (2015). Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 652–660). Springer volume 9351.

Chan, H.-P., Lo, S.-C. B., Sahiner, B., Lam, K. L., & Helvie, M. A. (1995). Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. *Medical Physics*, *22*, 1555–1567.

Chen, H., Dou, Q., Wang, X., Qin, J., & Heng, P.-A. (2016a). Mitosis detection in breast cancer histology images via deep cascaded networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 1160–1166). AAAI Press.

Chen, H., Wang, X., & Heng, P. A. (2016b). Automated mitosis detection with deep regression networks. In *13th IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1204–1207). IEEE.

Cheng, H.-D., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition*, *36*, 2967–2991.

Cireşan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012a). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems* (pp. 2843–2851).

Cireşan, D., Meier, U., & Schmidhuber, J. (2012b). Multi-column deep neural networks for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3642–3649).

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 411–418). Springer volume 8150.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Lawrence, T., & Prior, F. (2013). The cancer imaging archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, *26*, 1045–1057.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., & Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *SPIE Medical Imaging*. International Society for Optics and Photonics volume 9041.

CUDA (2017). What is CUDA?

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8609–8613).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (pp. 248–255).

Dhungel, N., Carneiro, G., & Bradley, A. P. (2015). Automated mass detection in mammograms using cascaded deep learning and random forests. In *IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–8).

Dhungel, N., Carneiro, G., & Bradley, A. P. (2016). The automated learning of deep features for breast mass classification from mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 106–114). Springer volume 9901.

Digital-Mammography-DREAM-Challenge (2017). 1.2M USD crowdsourced contest aims to improve breast-cancer detection through deep machine learning.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, *31*, 198–211.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)* (pp. 647–655).

D'Orsi, C. J. (2013). *ACR BI-RADS Atlas: Breast Imaging Reporting And Data System*. American College of Radiology.

Dos Santos, C., Marshall, P., Torresan, R., Tinóis, E., Duarte, G., & Teixeira, S. (2016). Abstract p4-01-04: Immunohistochemical and histological features of mammographic dense and non-dense tissue in breast cancer patients. *Cancer Research*, *76*, P4–01.

Dubrovina, A., Kisilev, P., Ginsburg, B., Hashoul, S., & Kimmel, R. (2016). Computational mammography using deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, (pp. 1–5).

Dundar, M. M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O., & Gurcan, M. N. (2011). Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, *58*, 1977–1984.

Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., Yankaskas, B. C., Kerlikowske, K., Onega, T., Rosenberg, R. D., Sickles, E. A., & Buist, D. S. M. (2009). Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy. *Radiology*, *253*, 641–651.

Elston, C. W., & Ellis, I. (1991). Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, *19*, 403–410.

Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., DOrsi, C., Elmore, J. G., Barlow, W. E., & Consortium, B. C. S. (2011). Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer Institute.*, *103*, 1152–1161.

Fonseca, P., Mendoza, J., Wainer, J., Ferrer, J., Pinto, J., Guerrero, J., & Castaneda, B. (2015). Automatic breast density classification using a convolutional neural network architecture search procedure. In *SPIE Medical Imaging*. volume 9414.

Fotin, S. V., Yin, Y., Haldankar, H., Hoffmeister, J. W., & Periaswamy, S. (2016). Detection of soft tissue densities from digital breast tomosynthesis: Comparison of conventional and deep learning approaches. In *SPIE Medical Imaging*. International Society for Optics and Photonics volume 9785.

Gastounioti, A., Conant, E. F., & Kontos, D. (2016). Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Research*, *18*, 91–103.

Ghosh, K., Brandt, K. R., Reynolds, C., Scott, C. G., Pankratz, V., Riehle, D. L., Lingle, W. L., Odogwu, T., Radisky, D. C., Visscher, D. W., Ingle, J. N., Hartmann, L. C., & Vachon, C. M. (2012). Tissue composition of mammographically dense and non-dense breast tissue. *Breast Cancer Research and Treatment*, *131*, 267–275.

Giger, M. L. (2014). Medical imaging and computers in the diagnosis of breast cancer. In *SPIE, Photonic Innovations and Solutions for Complex Environments and Systems (PISCES) II*. International Society for Optics and Photonics volume 918908.

Giger, M. L., Karssemeijer, N., & Schnabel, J. A. (2013). Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual Review of Biomedical Engineering*, *15*, 327–357.

Giusti, A., Caccia, C., Cireşari, D. C., Schmidhuber, J., & Gambardella, L. M. (2014). A comparison of algorithms and humans for mitosis detection. In *IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (pp. 1360–1363). IEEE.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *14th International Conference on Artificial Intelligence and Statistics* (pp. 315–323). volume 15.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*, 1153–1159.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, *2*, 147–171.

Hamidinekoo, A., Suhail, Z., Qaiser, T., & Zwiggelaar, R. (2017). Investigating the effect of various augmentations on the input data fed to a convolutional

neural network for the task of mammographic mass classification. In *Annual Conference on Medical Image Understanding and Analysis* (pp. 398–409). Springer.

He, W., Juette, A., Denton, E. R., Oliver, A., Martí, R., & Zwiggelaar, R. (2015). A review on automatic mammographic density and parenchymal segmentation. *International Journal of Breast Cancer*, *2015*, Article ID: 276217.

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, W. P. (2001). The digital database for screening mammography. In *In Proceedings of the 5th International Workshop on Digital Mammography* (pp. 212–218). Medical Physics Publishing.

Holland, R., & Hendriks, J. (1994). Microcalcifications associated with ductal carcinoma in situ: mammographic-pathologic correlation. In *Seminars in Diagnostic Pathology* (pp. 181–192). volume 11.

Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, *3*, 034501.

ICPR2012 (2017). contest (International Conference on Pattern Recognition). Tsukuba, Japan.

Irshad, H., Veillard, A., Roux, L., & Racoceanu, D. (2014). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. *IEEE Reviews in Biomedical Engineering*, *7*, 97–114.

Jamieson, A. R., Drukker, K., & Giger, M. L. (2012). Breast image feature learning with adaptive deconvolutional networks. In *SPIE Medical Imaging*. volume 8315.

Janowczyk, A., Basavanhally, A., & Madabhushi, A. (2017). Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. *Computerized Medical Imaging and Graphics*, *57*, 50–61.

Janowczyk, A., Doyle, S., Gilmore, H., & Madabhushi, A. (2016). A resolution adaptive deep hierarchical (radhical) learning scheme applied to nuclear segmentation of digital pathology images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, (pp. 1–7).

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. In *Journal of Pathology Informatics*. Medknow Publications volume 7.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 675–678). ACM.

Jiao, Z., Gao, X., Wang, Y., & Li, J. (2016). A deep feature based framework for breast masses classification. *Neurocomputing*, *197*, 221–231.

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., & Karssemeijer, N. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, *35*, 1322–1331.

Kooi, T., Gubern-Merida, A., Mordang, J.-J., Mann, R., Pijnappel, R., Schuur, K., den Heeten, A., & Karssemeijer, N. (2016). A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In *International Workshop on Digital Mammography* (pp. 51–56). Springer volume 9699.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303–312.

Kopans, D. B. (1992). The positive predictive value of mammography. *American Journal of Roentgenology*, *158*, 521–526.

Kothari, S., Phan, J. H., Stokes, T. H., & Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, *20*, 1099–1108.

Kowal, M., Filipczuk, P., Obuchowicz, A., Korbicz, J., & Monczak, R. (2013). Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in Biology and Medicine*, *43*, 1563–1572.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images, .

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lamb, P. M., Perry, N. M., Vinnicombe, S. J., & Wells, C. A. (2000). Correlation between ultrasound characteristics, mammographic findings and histological grade in patients with invasive ductal carcinoma of the breast. *Clinical Radiology*, *55*, 40–44.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2169–2178). volume 2.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS),* (pp. 253–256).

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop. In *Neural Networks: Tricks of the Trade* (pp. 9–48). Springer.

Lévy, D., & Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. In *Computing Research Repository - arXiv.org*. volume abs/1612.00542.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60 – 88.

Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, *6*, 26286.

Lopez, M. G., Posada, N., Moura, D. C., Pollán, R. R., Valiente, J. M. F., Ortega, C. S., Solar, M., Diaz-Herrero, G., Ramos, I., Loureiro, J., Fernandes, T. C., & Ferreira de Araujo, B. M. (2012). BCDR: a breast cancer digital repository. In *15th International Conference on Experimental Mechanics*.

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, *33*, 170–175.

Malon, C. D., & Cosatto, E. (2013). Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of Pathology Informatics*, *4*, 9.

Matheus, B. R. N., & Schiabel, H. (2011). Online mammographic images database for development and comparison of cad schemes. *Journal of Digital Imaging*, *24*, 500–506.

Medsker, L., & Jain, L. C. (1999). *Recurrent Neural Networks, Design and Applications*. CRC press.

MIT-Technology-Review (2017). 10 Breakthrough Technologies in 2013.

MITOS-ATYPIA-14 (2016). The International Conference for Pattern Recognition (ICPR), Detection of mitosis and evaluation of nuclear atypia score in Breast Cancer Histological Images.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, *19*, 236–248.

Muhimmah, I., Oliver, A., Denton, E. R., Pont, J., Pérez, E., & Zwiggelaar, R. (2006). Comparison between Wolfe, Boyd, BI-RADS and Tabár based mammographic risk assessment. *Lecture Notes in Computer Science*, *4046*, 407.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 807–814).

National-Health-Service (2016). Breast screening: professional guidance.

Neal, L., Tortorelli, C. L., & Nassar, A. (2010). Clinician's guide to imaging and pathologic findings in benign breast disease. In *Mayo Clinic Proceedings* (pp. 274–279). volume 85.

Ng, A. (2011). Sparse autoencoder. *CS294A Lecture Notes in: Stanford University*, *72*, 1–19.

Oliver, A., Freixenet, J., Marti, J., Pérez, E., Pont, J., Denton, E. R., & Zwiggelaar, R. (2010). A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, *14*, 87–110.

Oliver, A., Freixenet, J., Martí, R., & Zwiggelaar, R. (2006). A comparison of breast tissue classification techniques. In *International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI* (pp. 872–879). Springer volume 4191.

OpenCL (2017). The open standard for parallel programming of heterogeneous systems.

Pang, J.-M. B., Byrne, D. J., Takano, E. A., Jene, N., Petelin, L., McKinley, J., Poliness, C., Saunders, C., Taylor, D., Mitchell, G., & Fox, S. B. (2015). Breast tissue composition and immunophenotype and its relationship with mammographic density in women at high risk of breast cancer. *PloS One*, *10*, e0128861.

Petersen, K., Chernoff, K., Nielsen, M., & Ng, A. Y. (2012). Breast density scoring with multiscale denoising autoencoders. In *Sparse Methods for Signal Reconstruction and Medical Image Analysis Workshop at MICCAI*.

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, *5*, e1000579.

Rangayyan, R. M., Ayres, F. J., & Desautels, J. L. (2007). A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, *344*, 312–348.

Ranzato, M., Poultney, C., Chopra, S., & Cun, Y. L. (2006). Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems* (pp. 1137–1144).

Romo-Bucheli, D., Janowczyk, A., Romero, E., Gilmore, H., & Madabhushi, A. (2016). Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images. In *SPIE Medical Imaging* (pp. 979106–979106). International Society for Optics and Photonics.

Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M. A., Adler, D. D., & Goodsitt, M. M. (1996). Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging*, *15*, 598–610.

Salakhutdinov, R., & Hinton, G. E. (2009). Deep Boltzmann Machines. In *in Proc. of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 448–455). volume 5.

Samala, R. K., Chan, H.-P., Hadjiiski, L., Helvie, M. A., Wei, J., & Cha, K. (2016a). Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics*, *43*, 6654–6666.

Samala, R. K., Chan, H.-P., Hadjiiski, L. M., Cha, K., & Helvie, M. A. (2016b). Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis. In *SPIE Medical Imaging 9785* (pp. 1–7). International Society for Optics and Photonics volume 9785.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Shin, H.-C., Lu, L., Kim, L., Seff, A., Yao, J., & Summers, R. M. (2015). Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1090–1099).

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*, 1285–1298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations, arXiv preprint arXiv:1409.1556*, .

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stavros, A. T., Thickman, D., Rapp, C. L., Dennis, M. A., Parker, S. H., & Sisney, G. A. (1995). Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology*, *196*, 123–134.

Stewart, B. W., & Kleihues, P. (2014). *World Cancer Report*. Lyon, France: IARCPress, International Agency for Research on Cancer, WHO.

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S. et al. (2015). Mammographic Image Analysis Society (MIAS) database v1. 21, .

Sun, W., Tseng, T.-L. B., Zhang, J., & Qian, W. (2016). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, .

Sun, X., Sandhu, R., Figueroa, J. D., Gierach, G. L., Sherman, M. E., & Troester, M. A. (2014). Benign breast tissue composition in breast cancer patients: association with risk factors, clinical variables, and gene expression. *Cancer Epidemiology Biomarkers and Prevention*, *23*, 2810–2818.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Tabár, L., & Dean, P. B. (2005). *Breast Cancer-The Art and Science of Early Detection with Mammography*. ISBN: 3-13-131 371-6: New York: Thieme.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*, 1299–1312.

Tot, T., & Tabár, L. (2011). The role of radiological–pathological correlation in diagnosing early breast cancer: the pathologists perspective. *Virchows Archiv*, *458*, 125–131.

TUPAC16 (2016). Tumor Proliferation Assessment Challenge.

UK-Breast-Cancer (2016). UK Breast Cancer Research Symposium.

Van Diest, P., Van Der Wall, E., & Baak, J. (2004). Prognostic value of proliferation in invasive breast cancer: a review. *Journal of Clinical Pathology*, *57*, 675–681.

Veillard, A., Kulikova, M. S., & Racoceanu, D. (2013). Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information. *Diagnostic Pathology*, *8*, 1–3.

Veta, M., van Diest, P. J., Jiwa, M., Al-Janabi, S., & Pluim, J. P. (2016a). Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS One*, *11*, e0161286.

Veta, M., van Diest, P. J., & Pluim, J. P. (2016b). Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 632–639). Springer volume 9901.

Veta, M., Pluim, J. P., van Diest, P. J., & Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, *61*, 1400–1411.

Veta, M., Van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B., Vestergaard, J. S., Dahl, A. B., Cireşan, D. C., Schmidhuber, J., Giusti, A., Gambardella, L. M., Tek, F. B., Walter, T., Wang, C.-W., Kondo, S., Matuszewski, B. J., Precioso, F., Snell, V., Kittler, J., de Campos, T. E., Khan, A. M., Rajpoot, N. M., Arkoumani, E., Lacle, M. M., Viergever, M. A., & Pluim, J. P. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, *20*, 237–248.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016a). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, .

Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., & Madabhushi, A. (2014a). Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. In *SPIE Medical Imaging* (p. 90410B). International Society for Optics and Photonics volume 9041.

Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., & Madabhushi, A. (2014b). Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, *1*, 034003.

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016b). Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific Reports: PMC*, *6*, 27327.

Wolfe, J. N. (1976). Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, *126*, 1130–1137.

Xie, Y., Xing, F., Kong, X., Su, H., & Yang, L. (2015). Beyond classification: structured regression for robust cell detection using convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 358–365). Springer volume 9351.

Xing, F., Xie, Y., & Yang, L. (2016). An automatic learning-based framework for robust nucleus segmentation. *IEEE Transactions on Medical Imaging*, *35*, 550–566.

Xu, J., Luo, X., Wang, G., Gilmore, H., & Madabhushi, A. (2016a). A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, *191*, 214–223.

Xu, J., Xiang, L., Hang, R., & Wu, J. (2014). Stacked Sparse Autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology. In *IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (pp. 999–1002).

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., & Madabhushi, A. (2016b). Stacked Sparse Autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, *35*, 119–130.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *IEEE International Conference on Computer Vision* (pp. 2018–2025).